CONSUMER RESPONSE TO ROAD PRICING: OPERATIONAL AND DEMOGRAPHIC EFFECTS

A Dissertation
Presented to
The Academic Faculty

by

Adnan Sheikh

In Partial Fulfillment
of the Requirements for the
Ph.D. Degree in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
December 2015

CONSUMER RESPONSE TO ROAD PRICING: OPERATIONAL AND DEMOGRAPHIC EFFECTS

Approved by:

Dr. Randall Guensler, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Hunter
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Patricia Mokhtarian
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael Rodgers
Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Catherine Ross
School of City and Regional Planning
Georgia Institute of Technology

Date Approved: October 28, 2015

ACKNOWLEDGEMENTS

I would, first and foremost, like to thank my parents, Saba and Salman Sheikh, and my sister, Farah Sheikh, for their unending and unconditional support. I would not be here, or anywhere, without the love of my family. Thank you to my advisor, Dr. Randall Guensler, for encouraging me to pursue a doctorate and guiding me through every step of this process. Thank you to my dissertation committee for always giving me more to think about and keeping me honest. Thank you to my adventuring group, Aaron Greenwood, Stefanie Brodie, Maria Roell, Ricky Clousing, Jaime Shapiro, and Jesse Fortner, for letting me lead you into treacherous territory. Finally, thank you to Imran Shaukat, Jonathan Hess, Nicholas Czabaranek, John Gangler, and Patrick Kates for keeping me laughing while I was hundreds of miles away for the past five years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ACS	American Community Survey
AIC	Akaike Information Criterion
ARC	Atlanta Regional Commission
AVI	Automatic Vehicle Identification
FHWA	Federal Highway Administration
GP Lane	General Purpose Lane
GRTA	Georgia Regional Transportation Authority
GDOT	Georgia Department of Transportation
GPS	Global Positioning System
GTRI	Georgia Tech Research Institute
HOV Lane	High-Occupancy Vehicle Lane
HOT Lane	High-Occupancy Toll Lane
IIA	Independence of Irrelevant Alternatives
MFDU	Multi-Family Dwelling Unit
MSE	Mean Squared Error
OBU	On-Board Unit
RFID	Radio Frequency Identification
RTMS	Remote Traffic Microwave Sensor
SFDU	Single-Family Dwelling Unit
SR-91	State Route 91
SRTA	State Road and Tollway Authority
TCRP	Transit Cooperative Research Board
TRB	Transportation Research Board
UNT	Ubiquitous Network Tolling
VMT	Vehicle Miles Traveled
VOR	Value of Reliability
VTTS	Value of Travel Time Savings

SUMMARY

The High Occupancy Vehicle (HOV) lanes on Atlanta, Georgia's radial I-85 had long been providing sub-optimal throughput in the peak traffic hours, as the two-person occupancy requirement allowed the lanes to become heavily congested. The Georgia Department of Transportation converted 15.5 miles of HOV 2+ lanes to High Occupancy Toll (HOT) lanes, one in each direction on I-85. The lanes use dynamic value pricing to set toll levels based on the volume and average speed of traffic in the lanes. The goal of this research was to investigate the responses to toll lane pricing and the factors that appear to inform lane choice decisions, as well as examining values of travel time savings and toll price elasticity for users of the Express Lanes. This study of the metropolitan Atlanta I-85 Express Lanes operates at the microscopic level to examine the impact of demographic characteristics, congestion levels, and pricing on users' decisions to use or not use the I-85 Express Lanes.

After the introduction and literature review, the dissertation provides an overview of the data sources and the processing methods used to construct a usable analytical data set. The next chapter describes a major effort in the construction of this data set: that of pairing the lane use data with marketing demographic data. The following sections discuss the quality of the various data sources and the issues with them, as well as the opportunities for sample bias that arose as a result of the data processing and construction of the final data set.

The dissertation then proceeds in examining the value of travel time savings distributions as a whole and across different income segments and trip lengths. The

differences in these distributions among lower, medium, and higher income households were marginal at best. The results did not indicate that higher income households had the highest value of travel time savings results, as may have been expected. More substantial variation was found among trips of differing lengths within the Express Lane corridor. The modeling work discussed next provided a number of insights into toll lane use. The determinants of lane choice decision-making in the morning peak had notable differences from the determinants of the afternoon peak. The initial analysis involved models which were estimated across three different income segments to examine differences in decision making between low, medium, and higher income households. The results indicated that the parameters were largely consistent across the three segments. Further segmenting the households showed that lane choice determinants varied more within the 'Higher' income segment than across the original three-segment structure. In particular, the five-segment models illustrated lower elasticities with regard to corridor segment counts and toll levels for the highest-income households in the sample, as well as higher household income level elasticities for afternoon trips by that same cohort.

This research was among the first in the available literature to use revealed preference lane use data for both the toll lane users and the unpriced general purpose lane users. The use of household level marketing data, rather than census or survey data, was another unique characteristic of this research. The analysis of value of travel time savings with a demographic component that looks at household income has not yet been seen in the literature; similarly, the findings regarding differing behavior among very high income households appear to be unseen in the existing literature. The results from this analysis, such as willingness-to-pay values for different population segments, will be

useful inputs to the decisions surrounding future HOT implementations in the Atlanta region. The use of new data sources, the evaluation of those types of data sources, and the application of methods that have previously been unused in this field make up the primary contributions of this dissertation.

CHAPTER 1

INTRODUCTION

The concept of road pricing has been widely promoted by economists since Arthur Pigou first proposed the idea in 1920 (Pigou, 1920). Making users pay when and where they drive allows them to realize more of the external costs that they impose on others. Varying these tolls with traffic levels also has the potential to reduce congestion by managing demand. For decades, however, toll implementations involved predetermined variations in toll levels, falling short of the dynamically priced ideal. Now, with the ubiquity and affordability of technologies such as radio-frequency identification (RFID) short range transponders, more dynamic and economically efficient systems can be deployed.

In Georgia, the High Occupancy Vehicle (HOV) lanes on Atlanta's radial I-85 had long been providing sub-optimal throughput in the peak traffic hours, as the two-person occupancy requirement allowed the lanes to become heavily congested (Guin, 2008). The Georgia Department of Transportation (GDOT) sought to address this problem by converting the lanes to High Occupancy Toll (HOT) facilities (*HOV Strategic Implementation Plan Atlanta Region*, 2003). These HOT lanes restrict traffic to carpools with three occupants and to users willing to pay a toll, with the goal of maintaining free-flow conditions through pricing that changes based on lane conditions. The HOV-to-HOT project converted 15.5 miles of HOV 2+ lanes to HOT lanes, one in each direction on I-85. The HOT length begins at the junction with I-285, which forms a perimeter around Atlanta, and continues north into the surrounding suburbs. The lanes use dynamic

value pricing to set toll levels based on the volume and average speed of traffic in the lanes. GDOT's goal is to consistently achieve speeds of 45 miles per hour in the I-85 Express Lanes, and the dynamic pricing algorithms are designed to reflect this. The lanes have multiple entry and exit points, and the tolls are assessed using vehicle transponders attached to windshields and RFID tag readers located over the lanes on the freeway. Vehicles with occupancies of three or more travel for free in the HOT lanes and must also carry transponders. The toll lanes opened in October, 2011. Prices are adjusted at five-minute intervals for the various entry-and-exit trip combinations. Today, tolls range from $0.01 per mile in the off-peak periods to over $0.50 per mile in the peak hours ($0.16 to $11 for a complete traverse of the facility).

HOT lanes differ from other pricing schemes in that they offer users the choice to pay for improved service. Unlike cordon pricing systems, such as London's Congestion Charging Zone, or bridge and tunnel tolls, drivers may still use an adjacent free alternative without changing their route or mode. This means that corridor users make different decisions with every trip, including whether to use the priced or unpriced lanes. Drivers then choose the length they want to travel in the lanes. This is in addition to the prior decisions of whether to obtain a Peach Pass transponder and whether to take a trip in carpool mode or in toll mode. As such, each trip along the I-85 corridor now involves multiple decisions relating to use of the Express Lanes and/or the General Purpose lanes. This study will work at the microscopic level to examine the impact of demographic characteristics, congestion levels, and pricing on users' decisions to use or not use the the I-85 Express Lanes.

The I-85 HOT corridor in Atlanta is relatively unique in that RFID tag reads are taken from the toll lanes as well as from the unpriced General Purpose (GP) lanes. This means that this project can assess a user's choice to use, or not use, the lanes as a function of price and traffic conditions. Using privately sourced demographic data, this dissertation will model individual users' choices as a function of demographics, toll price, and operating conditions. The results will illuminate differences in Express Lane decision making behavior among different segments of the population.

The data used in this analysis are new and unique in a number of ways. The Express Lanes system provides disaggregated transponder detections in both the HOT and the GP lanes, a feature that for much of its operation was unique to this facility. As mentioned above, this allows researchers to know when a specific vehicle chose to use or to not use the Express Lanes. Use data also allows for direct comparisons of measures such as travel time and travel time variability between the priced and unpriced lanes using the same data source. Two other features of the data are unique to this implementation: the existence of partial corridor trips, which make up a majority of the trips taken, and the presence of repeat user data. For each transponder in the data set, records for all of that transponder's trips in both lane types are available. The available data also include trip lengths, toll amounts, start and end times, and whether the trip was in carpool or toll mode. These elements are quite rare, though not wholly unique, in HOT lane studies. The study also makes use of household-level socioeconomic data sourced from a marketing company. This is another data source that has not been used in toll lane or other pricing literature until now. This household level data provides an

alternative to the aggregated census sources and the costly and self-reported surveys that are commonly used for demographic information.

The goal of this research is to investigate the responses to toll lane pricing and the factors that appear to inform lane choice decisions. In addition, this dissertation will examine values of travel time savings and toll price elasticity for users of the Express Lanes. While these are common analyses in pricing projects, the dataset described in the following sections illustrates the uniqueness of this study. The dissertation will start with an overview of the available data and the data processing methods, followed by a comprehensive choice modeling analysis, and then an examination of the value of travel time savings and disaggregate demand elasticities. The existence of user history data allows for analysis using panel data methods to reduce bias in models of user response. These results can be used to inform discussions of the impacts of future projects on different demographic groups, and will allow for data-driven decision making to assess and minimize negative effects on different populations.

In terms of the significance of this dissertation, the ability to assess users' responses to congestion pricing as a function of user, system and pricing attributes is a novel use of a unique dataset, and an important input to policy decisions concerning future HOT lane investments and developments. This is especially significant as the Atlanta metropolitan region is considering spending more than $16 billion on a network of managed lane facilities (Atlanta Region Managed Lane System Plan, 2010). The results from this analysis, such as willingness-to-pay values for different population segments, will be useful inputs to the decisions surrounding future HOT implementations in the Atlanta region. This research can also be used in responding to equity and social

justice concerns and in future managed lanes toll and revenue estimations. HOT lanes are often given the moniker of "Lexus Lanes" due to the perception that they are used only by the rich (Patterson & Levinson, 2008). These perceptions may be based on limited studies that report very general results, such as the study by the Southern Environmental Law Center that used zip code-level data to report average incomes of Express Lane users (Atlanta Journal-Constitution, 2013). This analysis aims to provide a more accurate illustration of the ways that different income groups choose to use the facility. The modeling tools and methods that result will be transferable to other cities with similar data and toll lane infrastructure.

This study of the metropolitan Atlanta I-85 Express Lanes employs trip characteristics, facility operating conditions, and household demographics to provide a comprehensive overview of Express Lane users and their decision-making processes. The results provide the basis for a demand-modeling tool that can examine the response of consumers to different toll levels as a function of facility operating conditions and user demographics for forthcoming Express Lane implementations.

This dissertation begins with a background discussion and overview of the I-85 Express Lanes. The next section is a review of the existing literature concerning HOT lanes and other managed lane pricing implementations. The literature review describes the current methods and data sources used in pricing research and discusses some shortcomings of existing studies. The following chapter details the data sources and provides an overview of the data. This includes a thorough examination of the different operational data streams and the household demographic data used in the dissertation. Chapter four describes the data processing methods that converted the raw data to usable

formats and provides an overview of the constructed trips that make up the foundation of the data set. Chapter five describes the process of pairing the lane use data with the demographic data and examines the demographics of the resulting sample. The next chapter describes issues with the quality of the data and the attempts to address those issues. Chapter seven examines the potential avenues of bias in the data set. A preliminary investigation of modeling HOT lane use decisions at the trip level comprises the following chapter, followed by a comparison in chapter nine of the full data set with the more restricted data set used in the analyses . The dissertation continues in chapter ten by presenting the initial value of travel time savings analyses along with a more complete expansion of that work, including demographic and trip length factors. Chapter eleven uses hierarchical tree based regression methods to more closely examine the available variables. Chapter twelve expands on the initial modeling work with new variables and methods, and also examines demand elasticity among income groups and at different toll levels. The final chapter provides a conclusion to the research, summarizing the research findings, contributions, and the limitations of the study, and presenting potential extensions of the work.

Project Background

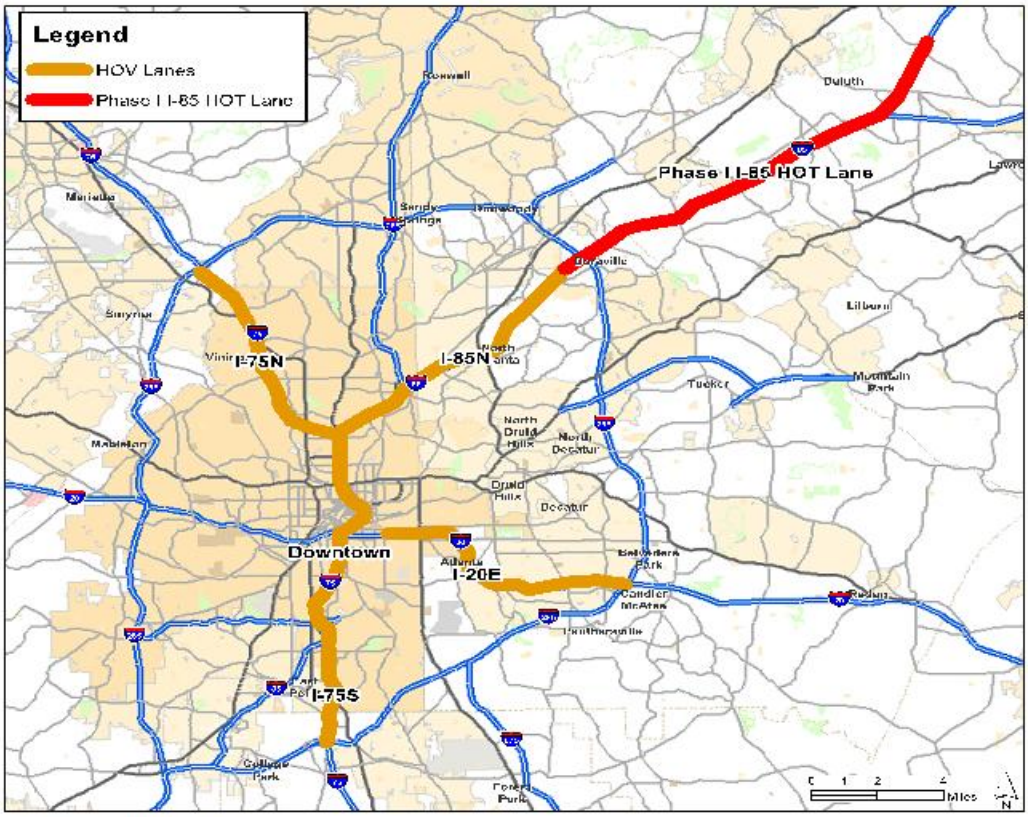
While roadway tolling has long been a common feature of America's infrastructure, dynamically priced tolls have only started appearing relatively recently. Orange County, California's State Route 91 (SR-91), the first fully automated, privately operated toll facility in the U.S., opened in 1995 (Sullivan, 2002). While the tolls on SR-91 are time-based, rather than congestion-based, later facilities implemented true value pricing. Such pricing schemes have been implemented in Minneapolis, Minnesota, San

Diego, California, Houston, Texas, Denver, Colorado, Seattle, Washington, and Miami, Florida (FHWA, 2012). The city of Atlanta, Georgia, is one of the latest cities to implement a dynamically-priced, fully automated pricing system. Atlanta's congestion levels have been rated among the worst in the country (Texas Transport Institute, 2012). In 2010, the USDOT awarded the Georgia Department of Transportation (GDOT), Georgia's State Road and Tollway Authority (SRTA), and the Georgia Regional Transportation Authority (GRTA) a $110 million Congestion Reduction Demonstration Program grant to convert underutilized HOV lanes into valued-priced HOT lanes. The grant also dedicated funds for increased bus service and improved park and ride lots along the corridor (Georgia Department of Transportation, 2013).

On October 1, 2011, the City of Atlanta, Georgia opened its first HOT lanes on the I-85 radial freeway. The Georgia Department of Transportation's (GDOT) HOV-to-HOT project converted almost 16 miles of HOV 2+ carpool lanes into HOT lanes, one in each direction. The HOT lane corridor begins at the junction with I-285, which forms a perimeter around Atlanta, and continues north into the surrounding suburbs. The State Road and Tollway Authority (SRTA), the operating agency, sets toll levels based on traffic volumes and average speeds of traffic on the corridor. SRTA's goal is to consistently achieve a speed of forty-five miles per hour in the Express Lane, and sets toll prices to manage demand for use of the HOT lane. The lanes have multiple entry and exit points, and the tolls are paid via electronic vehicle transponders known as Peach Passes. Prices are adjusted at five-minute intervals for the various entry-and-exit trip combinations. Vehicles with occupancies of three or more may travel for free in the HOT lanes, as may emergency vehicles, alternative fuel vehicles, and motorcycles,

provided that they register with the agency and carry Peach Pass transponders to use the Express Lanes.

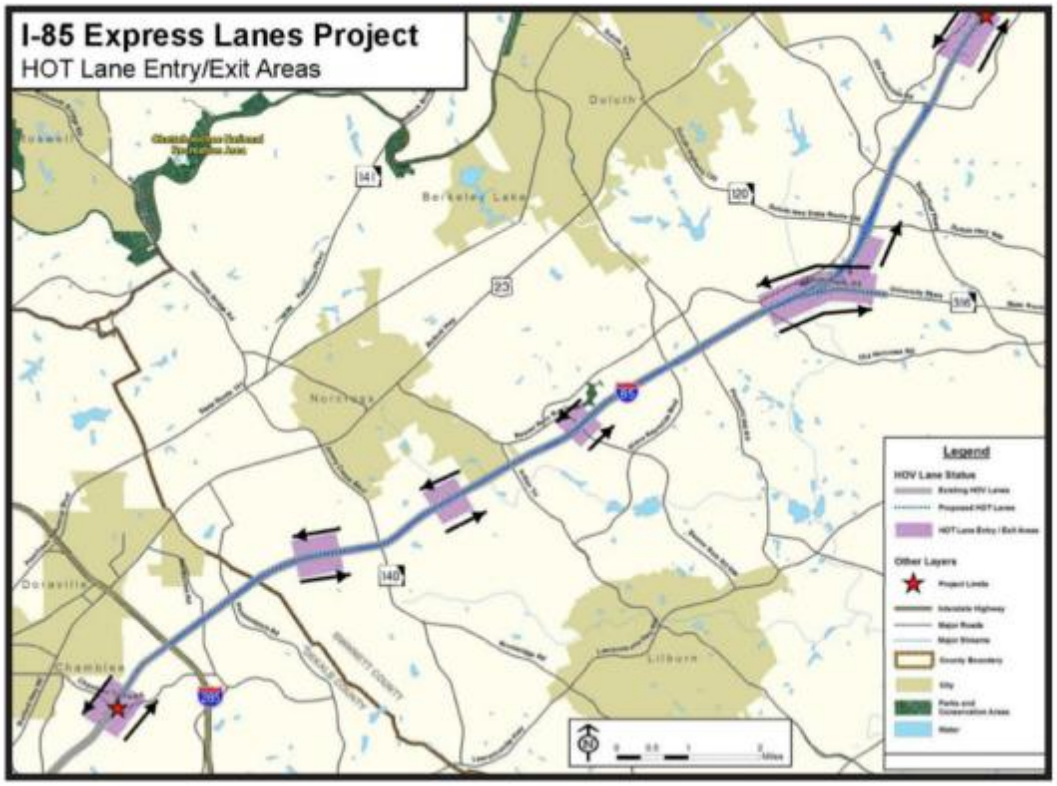
The I-85 Express Lanes stretch 15.5 miles from Chamblee Tucker Road (south of I-285) to both Old Peachtree Road and State Route 316 in the north. Figure 1 illustrates the length of the HOT lane corridor relative to the I-285 and the Atlanta metropolitan area. The lanes are equipped with automatic vehicle identification (AVI) scanners to read the Peach Pass RFID transponders. Thirty-five HOT gantries with AVI tag readers (RFID tag readers) sit above the Express Lanes. In addition, thirteen scanners (seven northbound and six southbound) sit above the general purpose lanes to detect transponders on general purpose lanes. The lanes are also flanked by ten enforcement cameras that capture license plates of vehicles (to positively identify vehicles without Peach Passes). Ten toll rate signs line the corridor and display the current toll rate for different trip lengths.



Source: GDOT, 2011

Figure 1: I-85 Express Lanes (Georgia Department of Transportation)

The Express Lanes are divided into six segments, named after the interchange closest to their entry points. The interchanges are Interstate 285 (285), Jimmy Carter Boulevard (JC), Indian Trail-Lilburn Road (IT), Pleasant Hill Road (PH), Old Peachtree Road (OP), and State Route 316 (SR316). These segments range in length from 1.76 miles to 3.60 miles. Lane access is provided by five dashed-line ingress/egress sections in the southbound and northbound directions. Those sections appear in purple below in Figure 2; sections vary in length from 0.35 miles to 0.66 miles. The RFID scanners placed on gantries over the HOT and General Purpose lanes provide a great deal of vehicle detection and lane condition data. Those data are described in Chapter 3.



Source: GDOT, 2012

Figure 2: I-85 Express Lanes Weave Zones

Research Framework

The research began with the reception and storage of the I-85 Express Lanes data from the State Road and Tollway Authority. These data were converted to the MySQL database format and stored on a secure server at the Georgia Institute of Technology. Georgia Tech researchers used these data to begin a preliminary investigation into the value of travel time savings exhibited by users of the Express Lanes (Sheikh, 2014). This investigation compared the amount of time toll lane users saved with the toll amount they paid and presented the resulting distributions in terms of dollars per hour. This initial work was followed by the extraction of license plate records from the SRTA data, which were matched in the motor vehicle registration database in a blind process to household address. This matching process allowed for the connection of SRTA vehicle data with

household-level socioeconomic data (Epsilon Targeting, 2013). These demographic data were obtained from a private marketing firm and were used in all subsequent research.

The research reported in this dissertation began with an investigation into the potential bias in the sample of households for which both Express Lane use and socioeconomic data were available. From here, researchers conducted a choice modeling analysis using the socioeconomic data in conjunction with corridor condition data and user history data. This analysis examined the determinants of lane choice decision making for users of different demographic segments under different traffic conditions. The preliminary value of travel time savings analysis was then expanded through the use of a greater scope of trips and the addition of a demographic component in the form of income segmentation. The final step was an investigation into price demand elasticity of users on the Express Lanes under different conditions and among different socioeconomic groups.

Research Contribution

This document is a PhD dissertation in partial fulfillment of the requirements for the PhD degree in the department of Civil and Environmental Engineering at the Georgia Institute of Technology. The dissertation makes a number of contributions to the body of research concerning road pricing in general, and High Occupancy Toll lanes specifically. The first of these contributions is the provision of new modeling results derived from new and novel data sources. The availability of automated detection data from the unpriced General Purpose lanes and the use of privately sourced household demographic data are rare, if not unique, in the realm of HOT lane use research. The dissertation also implements modeling methods that have not yet been applied to HOT lane research, such

as panel data methods for repeat observations by individual users. This dissertation work will also provide a basis for a spreadsheet-based demand modeling tool that may be suitable for future HOT implementation. Finally, the research involves the development of a host of data processing and modeling scripts that serve to: construct trips from disaggregated vehicle detections, estimate corridor conditions such as travel speeds and travel time reliability, and pair trip records with account, toll, and demographic data to provide a comprehensive overview of user characteristics and operating conditions at the time of each individual trip.

CHAPTER 2

LITERATURE REVIEW

The first HOT lane facility in the US was State Route 91 in California; operations began in 1995 (Sullivan, 2002). The unique characteristics of HOT lanes, including the presence of adjacent tolled and un-tolled alternatives, make them suitable for many operational and economic analyses. With the opening of SR91, impact assessment studies began appearing in the literature. Successive research has evolved since the SR-91 study. The research described in this dissertation includes conventional concepts and methods typically employed in studying HOT lanes, such as value of time, price elasticity of demand, and discrete choice modeling. The following literature review provides an overview of those concepts and of recent studies by other researchers. The review process pointed to a number of shortcomings and gaps in recent research that this dissertation hopes to rectify. The literature review begins with an overview of congestion pricing and case studies that illustrate the various forms it can take. The next section presents a discussion of research concerning the value of travel time and reliability. The literature review then describes the concept of price elasticity, its applications in transportation research, and presents a selection of relevant research. The last section of the review discusses previous choice modeling studies of HOT lanes with a focus on the different data sources (stated preference, revealed preference, and combinations of the two) and modeling methods that they have employed.

Congestion Pricing Overview

The lack of efficient road pricing has long been derided by economists. Randall Pozdena (2010), in his road pricing primer for the Puget Sound Regional Council, provides an overview of the arguments for pricing and the various forms it can take. Most infrastructure funding is currently raised by flat fees such as fuel taxes and registration fees. Funding does not vary by roadway or condition and is thus economically inefficient. In short, the driver is not paying the full cost of the burden he or she imposes on other drivers and the roadway. In describing the economist's position, Pozdena argues that "prices should…reflect the short-run marginal cost burdens imposed by the motorist." This includes both congestion and wear and tear on the facility. In addition, investment decisions are often made by political and level-of-service determinations, not cost/benefit analyses. This results in "the poor state of repair of road surfaces and bridges, and the dissipation of valuable time, fuel and capital resources due to congestion," along with resentment by users who do not see a connection between the fees they pay and the investments they enable (Pozdena, 2010).

Many of the traditional obstacles to road pricing are much more manageable today. Technological advances allow for transponders on highway gantries and in cars, or with on-board metering (GPS); hence, pricing no longer requires toll plazas on every road. Cost/benefit studies can identify areas in which pricing schemes would pay for themselves, reducing or avoiding the need for subsidy. One additional argument that Pozdena (2010) makes is that land use regulations and transit subsidies would be unnecessary if roads were properly priced. He argues that current policies are

economically inefficient, and that the same goals would be achieved in a less costly manner through pricing schemes (Pozdena, 2010).

Pozdena then describes different pricing systems with varying levels of complexity. The first is ubiquitous network tolling (UNT): Variable tolls are placed on freeways and arterials; however users must use on-board units (OBUs) which capture all travel. The next is freeway-only tolling: this can use gantry and transponder techniques, however users can divert to potentially less-used arterials to avoid the priced facilities. Area pricing, also known as cordon pricing, levies a toll as vehicles enter tolled zone. But, Pozdena argues that cording pricing is a poor approximation of pricing the individual paths. Finally, Partial Pricing generally takes the form of HOT Lanes, in which drivers have adjacent priced and unpriced options. Pozdena's model of economic and vehicle miles traveled (VMT) benefits indicates that UNT performs the best (Pozdena).

Guo and Yang (2009) look at congestion pricing from a different theoretical perspective: how it can be implemented to be Pareto-efficient. That is, how pricing can make people better off, without making any participants worse off. The authors' model uses multiclass users to account for how people of different income levels would react to the introduction of road pricing. Guo and Yang find that a scheme can be Pareto-efficient, if the tolling strategy reduces total system cost. Under their assessment, travel time can increase slightly across certain origin-destination pairs while still being Pareto-efficient, but only if the appropriate share of revenues is refunded to users to adequately compensate them for their increased travel times. This also addresses an issue that many have with the concept of congestion pricing, which is that it prices low-income users off

the roads. The study indicates that revenue refunding has efficiency as well as equity benefits (Guo & Yang, 2009).

Congestion Pricing in Other Jurisdictions

A number of pricing schemes have gained worldwide recognition. The cordon pricing systems in London, Singapore, and Stockholm have all been the subjects of numerous studies since their inceptions. Of these, the system in downtown Singapore is the oldest; it started in 1975 as a manual process, with drivers buying tickets to display in their windows and 'enforcement personnel' watching for violators at the entry points to the pricing zone. The complexity of the scheme increased rapidly; at one point personnel had to monitor 16 different license types. The overall effect was that traffic dropped 31% by 1988, despite a 77% increase in vehicle population (Chin, 2010).

In 1998, the manual Singapore system was replaced with an electronic system. Units in each vehicle contained cards with stored values, from which the fee was deducted upon each entry into the priced zone. A centralized control center identified vehicles without cards or with insufficient amounts on the card and sent out bills based on license plates. Volume counts were examined every three months and rates were adjusted to achieve the desired amount of traffic. The study attributed the success of the system to its flexibility (in adding new regions and varying the price based on demand) and to the public relations campaign emphasizing the traffic-management, rather than revenue-generating, nature of the program (Chin, 2010).

The city of London also implemented a cordon pricing scheme in 2003 in its central business district. Central London was seen as a suitable candidate due to low road capacity, heavy demand, and availability of alternative modes. Before the pricing

scheme, private automobile trips made up approximately 12% of total peak period trips. The cordon scheme was enforced by video cameras, and the system managed an average of 110,000 users per day as of 2006. The first few months of the program saw this percentage drop to 10%; removing almost 20,000 vehicles per day. Average speeds within the zone increased from 8mph to 11mph (an increase of 37%). Peak congestion delays decreased by 30%, and bus congestion delay decreased by 50% (enhancing the transit user's experience). Bus and subway ridership increased 14% and 1% respectively. While the scheme has been considered effective, various shortcomings have been noted. For one, the fee is not based on distance, time, or road congestion; a flat fee is imposed for all users. More congested roads cost the same as less congested roads. The system also has "relatively high overhead costs." The subway system, which is the alternate mode of many drivers, is "crowded and unreliable," though revenues from the scheme are being used to improve transit. The plan met with opposition from the public at first, however the plan was quickly accepted after implementation; other regions soon wanted to be included. In general, drivers showed more price elasticity than expected, which resulted in less congestion but also less revenue (Litman, 2006).

Stockholm began its own pricing scheme in 2006. Like those of Singapore and Central London, it was a cordon system. The study by Hamilton took a different approach in that it discussed the project's cost of implementation rather than public acceptance or effectiveness. While the system did not actually exceed its budget, its nature as a contracted project, the tight deadlines it faced, and the political pressure it was under, resulted in higher costs to the government. Because a major political party wanted a successful system in place before the next election, the government allowed the

contractor to overstaff its call center and increased payments for achieving deadlines. In addition, a number of requirement changes throughout the process resulted in higher costs to the government. Hamilton's paper found that the focus on minimizing technical and political risks raised costs significantly (Hamilton, 2010).

The case studies discussed above all offered key lessons and insights for other pricing projects. The systems demonstrated that pricing alone does not reduce congestion; changes to land use, improvements to public transportation, updated parking policies, and even road improvements may also be necessary. Alternatives to the priced roads, such as transit or unpriced routes, must also be provided or improved upon. The research also demonstrates the varying levels of complexity and the accompanying tradeoff with privacy. A common theme was the initial skepticism of the public, followed by greater acceptance after implementation. This was greatly aided by educational efforts and transparency to emphasize the traffic management nature of the programs. The case studies and pricing overview provide the most value as instructional resources for future pricing schemes.

Price Elasticity of Demand in Transportation

A common area of study in road pricing is the subject of price elasticity. The term refers to a ratio of change, such that a change in one variable can predict a change in a related variable. An elasticity value of 0.2 for Variable X with respect to Variable Y indicates that for every 1% change in Variable Y, there is a 0.2% change in Variable X. Any value less than 1.0 is considered "inelastic;" likewise, any value greater than 1.0 is considered "elastic." The amount by which a change in price causes a change in behavior has long been of interest to researchers in various fields. In the transportation industry,

fluctuations in gasoline prices and transit fares have often been examined for their effects on the demand for these products. With the more recent advent of dynamic road pricing, another avenue of consumer response has been made available for analysis. Technological advances allow for highly detailed datasets documenting price and volume changes within toll lanes, and sensitivities under a host of different conditions can now be considered. This dissertation will examine demand sensitivities of users on the I-85 corridor, and this section of the literature review examines elasticity results from other projects for purposes of comparison.

Before discussing the studies that evaluate actual elasticity levels, this section provides some background on how they are calculated. Pratt (2003), in a widely-cited publication by the Transit Cooperative Research Program (TCRP), presented four different methods of estimating price elasticities. Those methods are the point elasticity, the shrinkage ratio, the midpoint arc elasticity, and the log arc elasticity. Extending Pratt's work, Han (2009) examined those techniques of elasticity estimation and identified methods that are more appropriate for various situations. Han noted that despite the importance of the concept of elasticity to the transportation field, as of yet there is no agreed-upon technique for determining a value. Han argued that two general categories of techniques exist: statistical models and primitive formulas. Statistical models take into account more factors relating to travel demand, while primitive formulas have less onerous data requirements. Han evaluated three of the primitive methods described by Pratt (the exception being point elasticity) by comparing their estimator bias, variance, and mean square errors (MSE). Equation 1, for the point elasticity, most closely resembles the fundamental definition of elasticity in the economics literature.

Pratt argues that it is difficult to use in practice, however, as it requires knowing the demand curve (from which the derivative is taken) relating price and quantity.

$$\eta_\rho = \frac{dQ}{dP} \times \frac{P}{Q}$$

Equation 1: Point Elasticity (Pratt, 2003)

Equation 2, and the following equations, all provide different methods of approximating the point elasticity. The log arc elasticity is, according to Pratt, the formulation that "most nearly approximates point elasticity." As a result, it is formula that is used throughout TCRP Report 95, "Traveler Response to Transportation System Changes." However, Han (2009) estimated that the log arc elasticity had the highest MSE result of the three methods examined.

$$\eta = \frac{\Delta \log Q}{\Delta \log P} = \frac{\log Q_2 - \log Q_1}{\log P_2 - \log P_1}$$

Equation 2: Log Arc Elasticity (Pratt, 2003)

The midpoint arc elasticity, shown below in Equation 3, may be used when one of the variables in the log arc elasticity equation is zero. That is, when the starting or ending quantity or price are equal to zero. Han (2009) suggested that the midpoint arc method be used in circumstances where demand increases as it had the lowest MSE value.

$$\eta = \frac{\Delta Q}{(Q_1 + Q_2)/2} \div \frac{\Delta P}{(P_1 + P_2)/2} = \frac{\Delta Q(P_1 + P_2)}{\Delta P(Q_1 + Q_2)} = \frac{(Q_2 - Q_1)(P_1 + P_2)}{(P_2 - P_1)(Q_1 + Q_2)}$$

Equation 3: Midpoint (Linear) Arc Elasticity (Pratt, 2003)

Equation 4 provides the shrinkage ratio elasticity formula, which is typically used in transit and road pricing studies. It is perhaps the easiest to understand from an intuitive perspective, as the formula is defined as the relative change in demand divided

by the relative change in price. According to Pratt, these are often labeled "approximate point elasticities." This formula, however, gives different results for equivalent changes in opposite directions (Pratt, 2003). Han (2009) labeled this method the "most efficient technique among the three" in cases where demand decreases.

$$\eta = \frac{\Delta Q / Q_1}{\Delta P / P_1} = \frac{\left(Q_2 - Q_1\right) / Q_1}{\left(P_2 - P_1\right) / P_1}$$

Equation 4: Shrinkage Ratio (Pratt, 2003)

An important consideration when using these methods is the possibility that additional factors that have been excluded from the model affect demand and/or price. Looking solely at those two values ignores other variables that are likely to affect transportation demand, such as employment and prices of alternative modes. For this reason, most researchers use more robust methods, such as multiple regression and choice modeling, to estimate elasticities while controlling other factors. These methods make up the second category of elasticity estimators (i.e. the statistical models) that Han described but did not examine.

Oum (1992) provided a review of different concepts of elasticities and the model specifications that yield these different types of elasticities. The paper began by describing the difference between ordinary and compensated demand elasticities: ordinary price elasticities measure "both the substitution and income effects of a price change," while compensated price elasticities measure "only the substitution effect of a price change." Compensated elasticity is not estimated, however, as "it is a function of utility, which is not directly observable" (Oum, 1992).

Oum then goes on to describe mode-choice elasticities and regular demand elasticities. Mode-choice studies are those "which examine shares of a fixed volume of

traffic among modes." The elasticities that are estimated in this situation reflect substitutions between modes but "aggregate mode-choice studies…do not take into account the effect of a price change on the aggregate volume of traffic." Estimating regular demand elasticities requires acknowledging these changes in aggregate volumes, and disaggregate studies can correct this. Specifically, disaggregate studies which "include in the users' choice set the option of not making the trip" can generate regular demand elasticities. In the absence of this non-traveler data, the resulting elasticities are mode-choice rather than regular demand. As the data set for the I-85 Express Lanes contains both HOT and GP trips by users with transponders, this dissertation will estimate regular demand elasticities (Oum, 1992).

A study by Goodwin (1992) examined demand elasticity of fuel consumption with respect to fuel price and transit use with respect to fare price. Goodwin's main insight, however, was the difference between short-term and long-term values. While the papers Goodwin examined showed elasticity values well under -0.5 for, he estimated that these values increased by 50-200% over the long term. Short term elasticity of fuel consumption with respect to price averaged -0.27, while the average of long term studies was -0.71. The elasticity of traffic levels with respect to fuel price also increased from -0.16 to -0.33 from the short term to the long term. Goodwin's conclusions were that changes in behavior occur over time since more options are available as time increases, and that these long term elasticities make prices a strong mover of behavior. These changes include less car use or the purchase of more efficient vehicles; on the transit side, individuals may eventually move closer to stations if prices are attractive enough (Goodwin, 1992).

Because Atlanta's HOT lanes opened in October of 2011, there are now likely enough data to examine long-term elasticities. The additional controls that would be required, however, are out of the scope of this dissertation. These controls include data for changes in housing, employment, vehicle type, fuel prices, and other complicating factors over time. Future studies could potentially use the same data described in this dissertation, along with those additional measures, to assess long-term elasticities. Note that the steadily increasing toll rates, as discussed throughout this dissertation, will complicate these proposed studies.

A complicating matter in sensitivity studies is the fact that elasticities can be affected by a wide variety of factors, making comparisons between cities or even among different facilities within cities difficult. Hirschman (1995), in a study of elasticity values across bridges and tolls in New York City, noted that "elasticities can vary dramatically according to mode, time of day, travel purpose, household income, and by the amount and direction of the price change." The study looked at six bridges and two tunnels connecting the five boroughs of New York City. The authors developed time-series multiple-regression models from twelve years' worth of data for the various tolls; toll level, which increased from $0.75 to $2.50, was the main independent variable. The authors also included employment, gasoline prices, vehicle registrations, transit fares, and seasonal variations as independent factors. The shrinkage ratio was used as a method of checking their regression results, not the primary method, as it does not take into account the other factors the authors considered, such as employment (Hirschman, 1995). The results showed very low elasticities, as the authors predicted. Almost all of the values were much less than 1.0 in value and negative. The maximum elasticity estimated for

automobiles was -0.50, while the median was -0.10. Elasticities were higher where free alternatives existed or where transit alternatives are convenient, such as in Brooklyn and the Bronx. The authors concluded from this study that small increases in toll fares would not decrease congestion along the bridges and tunnels, but a "steep and sudden increase" could accomplish this (Hirschman, 1995).

A more recent study of elasticity values, although one that focused on Spain rather than the U.S., was written by Matas and Raymond (2003). The authors assessed seventeen years' worth of data concerning Spain's toll roads with a focus on examining why elasticity values differ on different segments. The model the authors created estimated that corridors with high traffic volumes were generally inelastic, while those that had a good alternative road (with high speeds) were more elastic. If the number of heavy vehicles on those alternatives roads increased, however, the toll demand became less elastic. Longer roads resulted in higher elasticity, likely due to the higher total price to be paid. Tourist areas exhibited higher levels of inelasticity, and overall demand was estimated to be more responsive to GDP than to gasoline prices. Another notable point the authors made was that setting a toll too high may create too much demand on the alternative road, increasing maintenance costs and environmental impacts. Reducing such a toll may actually decrease the total cost of the infrastructure. The results overall confirmed literature that suggested that demand is generally more elastic where there are good, un-tolled alternatives (Matas & Raymond, 2003). This suggests that as the I-85 HOT lanes operate alongside free General Purpose lanes, elasticity values should be higher than they would be in the absence of a free alternative.

Low elasticity values such as those described in the studies above have long been used by politicians and planners to develop pricing and transportation policy. Litman (2010) looked at the policy-related implications of low versus high elasticity values and the structural factors that help define them. Litman referred to the "rebound effect," in which higher fuel prices lead to the purchase of more fuel efficient cars and thus a driver's VMT actually increases. The "rebound effect" may have a direct relationship with elasticity: if elasticity is low, rebound effects may be small. Low elasticity and rebound effects support fuel efficiency mandates (in the case where the goal is to reduce fuel consumption), since greater fuel efficiency would reduce consumption and not raise external costs such as "congestion, accident risk and sprawl." Higher rebound effects and elasticities may make fuel efficiency standards less effective at reducing VMT and instead argue for pricing schemes (Litman, 2010). Unfortunately, the paper does not address the previously mentioned phenomenon of varying elasticities among different cities or facility types and alternatives.

The historical evidence Litman cites demonstrates the decrease in elasticity values between 1960 and 2005. This is reflected in the studies discussed above: nearly all of the elasticity values are very inelastic. Litman argues that the 20th century saw increases in VMT due to higher vehicle ownership and numbers of driver's licenses, more women in the workforce, expanded highways, worsening transit service, and low-density development. All of these factors in turn reduced elasticity. Litman goes on to suggest that other changes may now be increasing elasticity. These include an aging population of retirees and the elderly, who commute less and have lower incomes. Stagnating incomes combined with increasing fuel prices will also likely increase elasticity.

Increasing investments in pedestrian, bicycle, and transit infrastructure, along with more traffic congestion, may also serve to increase price sensitivity by providing alternatives with acceptable levels of service. This is reflected in recent elasticity research that has shown an upward trend in sensitivities since 2005 (Litman, 2010).

As mentioned above, lower elasticity values favor fuel efficiency mandate increases, because directly increasing the cost of driving will not cause drivers to drive less over the short term. Higher sensitivity values, however, favor pricing schemes such as road pricing. Transportation analyses have typically used elasticity values which according to Litman are now too low. Studies by the USDOT, for example, underestimate elasticity and thus the benefits of pricing schemes (Litman, 2010).

The various elasticity studies paint a complex picture of the effects of congestion pricing systems. Pricing schemes, if implemented, should consider spillover effects and traffic diversion that may increase maintenance costs. In many instances, the most significant impacts will take many years to be realized as sensitivities are low over the short term. On the other hand, changing demographics and political concerns may herald a reversal of the trend of low elasticities. This has many implications for transportation and planning policy, as elasticity levels affect decisions regarding fuel prices, transit fares, pricing plans, land use policies, and more.

This issue of traffic diversion onto unpriced roads was the focus of a study by Swan and Belzer (2010). The paper sought to estimate elasticity in response to road tolls by examining truck data in Ohio. The data was used to estimate the amount of VMT diverted from Ohio's highways onto its secondary roads. The focus here was on the effects of high elasticity values caused by the availability of suitable alternatives: the null

hypothesis the authors used, that no traffic would be diverted to other roads, was negated. Sixteen of the thirty-three routes had "significant positive coefficients" for diverted VMT. The results showed that the quality of an alternative road would increase the elasticity of a tolled road; as a result, "relatively small toll increases can lead to significant diversions of traffic from highways to secondary roads." The authors used these results to make the case against setting toll rates beyond marginal costs and arguing that "there may be a role for the Federal government in regulating state or local toll rates where they interfere with interstate commerce" (Swan & Belzer, 2010).

In addition to road and toll elasticity, transit fare elasticities also receive a great deal of attention. Todd Litman of the Victoria Transport Policy Institute published a review of elasticity literature from 1991 to 2004 with a focus on cross-elasticities, defined as "the percentage change in the consumption of a good resulting from a price change in another, related good." Litman began by listing various factors that affect elasticities. User type, such as choice rider versus dependent rider, low versus high income, etc., played a role. Dependent transit riders had lower elasticities, and commute trips were less price-sensitive. Different types of price changes also resulted in different elasticity values: fare changes, service changes (since service affects non-monetary costs), and parking prices had higher elasticity values than other changes. The direction of the change also made a difference: users were more sensitive to fare increases than to decreases. Hence, elasticities were price direction dependent. Elasticity also varied by mode, since different modes (such as bus versus rail) serve different markets (Litman, 2004).

Litman then moved on to transit elasticity values themselves, arguing that older studies looked at short- and medium-term effects and neglected long-term elasticities, which are typically two to three times the size. This agrees with the argument put forward by Goodwin in his study. In addition to looking at sensitivity with respect to fare, Litman (2004) examined the effects of service changes on ridership. Cross-elasticities relating to transit ridership with respect to fuel price were also discussed. Table 1 below summarizes the elastic ranges that resulted from Litman's study. Most of the values are inelastic; only a few long-term sensitivity ranges reach or exceed 1 (Litman, 2004).

Table 1: Transit Elasticities and Cross-Elasticities (Litman, 2004)

	Market Segment	Short Term	Long Term
Transit Ridership WRT transit fares	Overall	-.2 to -.5	-.6 to -.9
Transit Ridership WRT transit fares	Peak	-.15 to -.3	-.4 to -.6
Transit Ridership WRT transit fares	Off-peak	-.3 to -.6	-.8 to -1.0
Transit Ridership WRT transit fares	Suburban Commuters	-.3 to -.6	-.8 to -1.0
Transit Ridership WRT transit service	Overall	.5 to .7	.7 to 1.1
Transit Ridership WRT auto operating costs	Overall	.05 to .15	.2 to .4
Automobile travel WRT transit costs	Overall	.03 to .1	.15 to .3

Values of Travel Time and Reliability

Value of time is an important concept in transportation modeling, with applications for every mode and a place in many different modeling applications. The amount of money that a user will pay to save some increment of time, typically expressed in dollars per hour, has implications for transportation policy, frequency, pricing, and for the distributional impacts of different planning decisions. High Occupancy Toll lanes provide a unique way of studying values of time for different populations, as the facilities pair a priced alternative with an adjacent free alternative. The two differ in price and typically in performance, allowing for direct comparisons of trips along the same route. Similarly, a great deal of research has investigated the value of travel time reliability, though methodological complications and the difficulty of measuring user perception make it harder to identify and isolate this result.

Some previous studies, such as those by Small (2005), Levinson (2011), Liu (2007), and He (2011), have used econometric methods to generate value of travel time estimates. These studies often involve stated preference survey data, or some combination of survey and revealed preference data. In some cases, the trip characteristics must be estimated or simulated. He (2011), for example, approached the issue of having only HOT-lane choice data by generating "simulated" choices to use the unpriced lanes; whenever a user was not seen in the managed lane, they were assumed to be in the free lanes (2011). This assumption is not necessary for this dissertation, as the lane detection data identify General Purpose lane trips as well as HOT lane trips. Small used travel time estimates from student field work as a factor in his model for California State Route 91 (2005). The mixed logit models in these studies allow for random, rather

than fixed, coefficients and thus the resulting values of travel time are presented as distributions. Small estimated a median value of time of $21.46/hour (2005), while Levinson estimated values of time ranging from $3.40/hour to $20.56/hour (2011). Liu reported VTTS results ranging from $6.82/hour to $27.66/hour (2007). Devarasetty (2012) estimated a value of travel time savings of $51/hour on the Katy Freeway in Texas.

Burris, et al. (2012) used a strictly revealed-preference approach, without choice modeling, by comparing HOT trip times in Minneapolis and San Diego to General Purpose travel times generated from loop detector data. Using revealed preference data, toll paid and travel time saved, Burris estimated median VTTS figures of $73/hour and $116/hour for the morning and afternoon peaks in Minneapolis, and $49/hour and $54/hour for similar periods in San Diego. These values were calculated from five-minute averages of HOT speeds and volumes, GP speeds and volumes, toll rates, and trip counts. That study also explained that HOT users can see GP conditions before making the choice to use the priced lane, and so they have some knowledge of the potential time savings. Other work by Devarasetty (2013) showed that HOT users actually overestimate their time savings by an average of 11 minutes, which has implications for revealed preference willingness-to-pay research based on HOT lanes. The study by Burris was unique in that it was one of the few to rely solely on automated revealed preference data.

In the absence of survey data, it is difficult, if not impossible to identify whether users are paying for travel time savings or reliability. The answer is likely that users are paying for some combination of the two benefits, as evidenced by other studies that do incorporate stated preference surveys (Devarasetty, et al., 2012). Travel time reliability is

also often studied with econometric methods involving both stated and revealed preference data. The study by Carrion-Madera and Levinson (2012) used survey results with GPS data by presenting users with fixed route choices and studying their preferences. The resulting values of reliability ranged from $0.68/hour to $18.23/hour in the Minneapolis-St. Paul region. Small estimated a median value of reliability (VOR) of $19.56/hour in his study of SR91 in the Los Angeles, California region (2005), while Devarasetty et al. used survey results to come to a combined VTTS and VOR figure of $50/hour. This dissertation examines the reliability benefits of the HOT lanes, but in the absence of survey data it cannot assign a value to those benefits.

The literature concerning values of travel time savings and reliability as it relates to High Occupancy Toll lanes differs greatly among the different facilities under examination. The results show high levels of variability in the estimated values of time and reliability, ranging from $3.40/hour to $116/hour. Data limitations and methodological differences, such as estimated travel times and imputed lane choice decisions, make it difficult to directly compare the results of different studies. Like demand elasticity, values of time may differ by location due to factors that are not captured in these models. Other factors, such as political considerations, which may limit toll amounts, or safety benefits of HOT lanes, which users may value in addition to the time savings, may also contribute to these differences.

Additional value of time research has focused on other modes and facilities, rather than on HOT lanes, but contains worthwhile insights for value of time research. These studies examine the interaction of value of time with income, the role of supply-side uncertainty in toll-setting, and the effects of different levels of driver information on the

impact of tolls. Among these studies is a paper by Mohring, et al. (1987), which attempts to estimate monetary values for travel time and waiting time as they relate to income. Such an effort has been historically difficult due to limited data sets. The authors used a "disutility" coefficient for waiting time, which quantifies the adverse effects caused by a process (such as travel). In the case of the study, the authors assigned a disutility range of 0.3 to 0.5 to travel time versus waiting time. This indicates that travelers assign a disutility value to time spent traveling of 30-50% of the value of time spent waiting. For example, if the disutility value of an hour waiting is $10, the disutility value of an hour spent traveling is $3-5. The study uses data from the Singapore Bus Service to determine whether users would rather take a more expensive and comfortable bus that was available immediately or wait for a cheaper option. A "comfort premium" was assigned to the more expensive and luxurious bus. The authors found that the comfort premium increases with trip distance and during peak periods. Peak period riders were found to have higher values of wait times and responded more to convenience factors. The results showed that an average rider with a 50 minute peak-period trip is only 19% likely to wait for a nicer bus of equal fare if another bus is immediately available, while a rider planning a short trip is 60% likely to wait for the cheaper option. The results of the paper show that, as expected, the ratio of waiting-time value to income increases with income. In addition, nonwage earners in households have lower waiting-time values than wage earners in the same households (Mohring, Schroeter, & Wiboonchutikula 1987).

Two other studies looked at complications in common congestion pricing models, including supply-side uncertainty and random capacity and demand. Boyles et al. (2010) argue that traditional calculations of the marginal costs of traveling do not account for

uncertainty, such as "incidents, weather conditions, [and] fluctuations in travel demand." The authors relate this to congestion pricing by asking whether tolls should vary in response to disruptions: If an accident occurs, should tolls increase to keep drivers off the road? Or should they not increase since drivers expect better conditions with higher tolls? The authors discuss responsive versus unresponsive tolls and find that to properly account for uncertainty, "unresponsive tolls must be set higher than responsive tolls" (Boyles, Kockelman, & Waller, 2010).

Lindsey (2009) also looks at driver information as it relates to tolls, investigating three scenarios: 1) users have perfect information about conditions and the toll reflects that perfect information; 2) users have imperfect information and tolls reflect that imperfect information; and 3) tolls are set using less information than the users have. The paper focuses on highways, where "capacity and demand shocks are common." It notes that much congestion is due to nonrecurring events; crashes account for most congestion in urban areas. Tolls give drivers information about the level of congestion in non-toll lanes, but a toll level may mean different things. The author concludes that in the first two cases, toll revenues will pay for "optimal capacity of a facility," but not in the third case. The paper focuses on highways, where "capacity and demand shocks are common." It notes that much congestion is due to nonrecurring events; crashes account for most congestion in urban areas. Tolls give drivers information about the level of congestion in non-toll lanes, but a toll level may mean different things: "tolls cannot convey complete information about the state" (Lindsey, 2009).

High Occupancy Toll Lane Decision Making Studies

With the growing popularity of High Occupancy Toll lanes in the US has come a corresponding amount of modeling research concerning these lanes. This literature is varied in both methods and data sources. The various studies use different types of choice and regression models, and the data may include revealed preference automated reporting, traveler surveys of trip and demographic characteristics, or some combination of the two. The results from these different studies are instructive, even when the methods or data differ from what is being investigated here. Different studies also point to different determinants as being significant in route or mode choice decisions; this may also be a function of differing data and methods. As a result, it is difficult to make direct comparisons between studies and to judge whether previous analyses are confounded.

Stated Preference Studies

Stated preference travel behavior modeling studies are very common. These studies measure a variety of characteristics and attitudes, including socioeconomic attributes and expected responses to potential situations. Asensio and Matas (2007) used survey results to examine the impacts of travel time variability and to put a value on travel time reliability. The stated preference method is used here to look at the value of reliability for different user and trip characteristics. The paper uses a mean-variance model in its analysis, with travel time variability represented by the standard deviation of travel time. The resulting models gave values of travel time savings and reliability, as expected; the stated preference method allowed for segmentation by the flexibility of a respondent's arrival time. This dissertation will examine travel time variability using

vehicle detection data, but arrival time flexibility is an example of the category of data that is not available in a strictly revealed-preference data set.

Burris and Pendyala (2002) estimated multinomial logit models for pricing participation and frequency of participation based on demographic surveys of a variably-priced facility (a bridge rather than an HOT lane in this case). The data were also revealed preference in that the respondents were users of the facility. Like the research undertaken in this dissertation, the Burris and Pendyala paper aimed to "describe the participation of travelers in variable pricing programs as a function of their socio-economic and commute characteristics." One shortcoming of the stated preference approach that the authors identified is that "travelers often tend to overstate their potential response to a hypothetical stimulus in stated preference surveys." Studies by Hensher (2001) and Calfee, Winston, and Stempski (2001), described below, detailed other shortcomings with the stated-preference approach: it is inappropriate for prediction, it may yield biased results depending on the design of the survey, and it may be more expensive or difficult to achieve large sample sizes.

Calfee, Winston, and Stempski (2001) used survey data to estimate value of automobile travel time for respondents in "major U.S. metropolitan areas." The authors estimated ordered probit, rank-ordered logit, and mixed logit models and then compared the resulting values of congested time from the different methods. The values ranged from $2.92/hour to $5.47/hour for the ordered probit, $3.12/hour to $5.47/hour for the ordered logit, and $3.17 to $5.47 (mean values) for the mixed logit. Calfee, Winston, and Stempski noted that stated preference studies must be designed with an "accurate ordering of preferences," and that other design decisions (such as ordinal versus cardinal

rankings) may cause bias. The sample size of this study was 1,170 respondents, and the authors made the point that with stated preference methods, it may be more expensive or difficult to get large samples.

Yan, Small, and Sullivan (2002) also used survey data, this time from State Route 91 in California, to estimate more complex joint and nested logit choice models. In this study, the route and mode choices were distinct, and additional models included transponder choice. Yan, Small, and Sullivan sought to investigate the effect of toll changes on "vehicle occupancy or time of day instead of or in addition to changing route." The value-of-time results estimated by the choice models were in the range of $13-16 per hour. This is much lower than the results that were estimated in the preliminary study of I-85 described later in this document. The authors note that these results were for congested travel time, "which is known to have a higher value than that of uncongested time." In addition to estimating value of time and elasticity results, the authors noted that it is "quite possible that pricing demonstrations in which there is a free road parallel to the priced road do not capture the full range of behavioral responses." This is because drivers do not need to consider more drastic behavioral changes if they can simply take the unpriced roadway parallel to the same route. Here the benefits of stated preference data include the ability to investigate more choices and to include individual-level factors such as sex. The downsides include potentially lower value of time results than would be seen in revealed preference studies.

Another study by Burris (2006) estimated multinomial logit models from stated preference surveys to explain HOT lane use. This study was unique in that it excluded mode characteristics, explaining travel time and required occupancy, for example, "were

felt to be implicitly included in the traveler's choices." These results indicated that users of the facility were "significantly more likely to be over 65 years old, have a post-graduate degree, have a household income greater than $100,000 per year, and be on a school-related trip." The paper illustrated that unique geographical aspects of a study, in this case the prevalence of schools near the end of the facility, may have a large impact on the results. The stated preference nature of this and other studies allows analysts to assess the influence of trip purpose, which generally cannot be derived from observational data alone. A frequently-cited study by Li (2001) estimated logistic regression models using similar methods to those outlined in this dissertation. The determinants were categorized as demographic characteristics, financial capability, and travel characteristics. Demographic characteristics included household size, household "type," gender, and age, while "financial capability" referred to household income. Travel characteristics included trip length, vehicle occupancy, a commute trip dummy variable, and trip frequency. The author hypothesized that the HOT lanes would be more frequently used for commute trips, long-distance trips, high-occupancy trips, and trips by frequent users, women, and larger households. The list of resulting significant variables, including age, financial ability, vehicle occupancy, would be interesting to compare with the results of the revealed preference study proposed here.

Hensher (2001) created mixed logit models from survey data and compared the value of travel time savings results with those from standard multinomial logit models. The mixed logit models "[produced] higher estimates of values of time savings compared to the multinomial logit model." Hensher estimated values of free flow, slowed down, and start/stop time with both MNL and Mixed Logit models. The resulting mean values

ranged from $0.06 to $5.90 higher for the Mixed Logit cases, with an average difference of $0.37 for the value of free flow time and $3.02 for the value of stop/start time. This supports the cases for including random coefficients where appropriate in this dissertation's choice models, as the resulting values of time may be significantly different. Hensher also discussed the importance of preference heterogeneity in transportation modeling, as neglecting this issue leads to serial correlation in the error term. Significantly, ignoring preference heterogeneity can also have "impact on the marginal rates of substitution between attributes." This issue may directly impact the elasticity studies in this dissertation. The study locations in question were seven cities in New Zealand. One final important note from this paper is its description of revealed preference data as "'dirty' from the point of view of statistical estimation," as there is "often too much confoundment in RP data." However, Hensher stressed the importance of revealed preference data in prediction, and stated that "the SC component of a data set is useful only in improving the statistical efficiency of the parameters associated with the design attributes."

The studies discussed above illustrate many of the benefits of stated preference studies: flexibility in choices and in scenarios presented to the respondents, the ability to control for confounding and correlated factors, and the ability to capture the order of a respondent's preferences. They also reveal some of the downsides in the form of potentially unrealistic responses and value of time estimates that do not match those from revealed preference studies, as well as potential sample size limitations due to expense and response rates. Other important points include the potential for results to be specific to a certain geography and some of the potential pitfalls of revealed preference data.

Stated and Revealed Preference Studies

Studies that combine both stated and revealed preference data are typically described as the most valuable, as they can capture the benefits of both types of data and make up for the shortcomings of each. For revealed preference data, these shortcomings include the fact that only the 'most preferred' option is reported, there may be correlation among the different variables, there may be a lack of variation in the data, and important factors may be excluded. In the case of this dissertation, variables representing traffic conditions may be correlated with toll amounts and time savings, for example.

Borjesson (2007) estimated mixed logit models for departure time choice using both stated preference and a form of revealed preference data. In this case, the revealed preference data was extracted from a model of the Stockholm network named "CONTRAM." While the travel times from the model were simulated, they were described by Borjesson as "actual mean travel times." Borjesson describes the benefits of combined RP and SP models, as in this case the revealed data are highly correlated. The author modeled departure time choice as a function of travel time variability, but "high correlation of mean travel time and travel time uncertainty in revealed preference data [made] accurate estimation of the trade-offs unfeasible." Borjesson later cites this "high correlation between mean travel time and travel time uncertainty in RP data" as a primary factor in the lack of travel time uncertainty studies using RP-only data. The paper does not discuss whether perceptions of travel time uncertainty affect the departure time decision, or whether actual uncertainty is the contributing factor. A related issue appears in the research proposed here, as values of toll amounts, travel time, and volumes are very likely to be correlated. However, the article ultimately concludes that stated

preference data is “less trustworthy for trip timing analysis and forecasting,” the goals of the paper.

Two of the most frequently cited revealed preference studies are those by Lam and Small (2001) and Small (2005). Even these studies, however, used both revealed preference data and stated preference data. Lam and Small (2001) used surveys asking for vehicle occupancy, job characteristics, and other information. In this highly-cited study, the average travel times for the models were estimated using a “standard engineering algorithm” and volume and vehicle density from loop detectors. The resulting travel time savings for the California State Route 91 lanes under examination were 5.9 minutes in 1998, a value that Lam and Small describe as small in magnitude and which “makes [their] results vulnerable to measurement error.” This value is within the same order of magnitude as the median travel time savings of the I-85 Express Lanes, relative to the entire corridor. An important note in this study, which also relates to other loop-detector based studies, is that there are “many assumptions required to convert loop detector data into speeds estimates” (Lam and Small, 2005). The dissertation research presented herein does not have to rely on estimated travel times, as the data include actual travel times in both HOT and General Purpose lanes.

The study by Lam and Small used binomial logit models for the route choice models, and included various measures of variability. The paper includes a discussion of the endogeneity in the models, namely in the option to switch to another route. Lam and Small then included time-of-day choice in their next models, but this came from the survey data, and hence cannot be repeated in this dissertation work. The authors also address the issue of their data covering a portion of the actual trip length by both ignoring

this limitation (so that its effects are in the alternative-specific constants) and by estimating those missing travel times. These two methods are both options for this dissertation. Finally, the authors examine transponder choice and find that "transponder installation has its own determinants, distinct from those of the daily decision of whether or not to use the transponder." As for the route choice results, the authors report that "work-hour flexibility [provided by surveys] and total trip distance seem to influence the daily decision of which route to take" (Lam and Small, 2005).

The Small (2005) paper estimated mixed logit models based on both revealed and stated preference data, with some important points for this dissertation. The author notes that revealed-preference studies "have been hampered by collinearity among cost and travel-time variables" and that "they have not accounted for heterogeneity in cost or travel time elasticities." An interesting point is that the author does not name any of these revealed-preference studies. Similarly, the revealed preference data used in the study is self-reported and comes from telephone surveys (Small 2005).

A number of other studies had important points that are relevant to this dissertation. Liu (2007) was a rare study that used revealed preference data, in the form of loop-detector data, to estimate mixed logit models of route choice. The main determinants in this study were "travel time, reliability, and cost." The study was also unique in that it examined values of travel time and reliability as they differed with departure time; that is, it did not assume them to be constant across the hours under study. Liu did not include demographics in that work; this dissertation will. Hess (2005) discussed mixed logit models with positive coefficients for travel time; these models indicate that users gain more utility from longer trips. The author notes that these are

typically seen as the result of model misspecifications or the lack of explanatory power in the data, and proposes other interpretations (Hess 2005). Goodall and Smith (2010) wrote a paper with some worthwhile methodological variations, such as separating "daily users" of the MnPASS HOT lanes from less frequent users in their models to achieve a much better fit. The paper concluded that "pricing has a negligible influence" on lane use because almost 90% of the facility users were daily users, and that drivers may "use the HOT lanes as insurance against unanticipated congestion." On the I-85 Express Lanes, however, only 3.5% (4231 out of 120582) of transponders used the priced facility more than 200 times in 2012 (four times a week for 50 weeks of the year). The authors also raise the question of what has the greater impact in lane use decisions: current conditions vs. previous experience.

A common theme in the literature concerning HOT choice modeling is the set of shortcomings of both stated and revealed preference data. Data sources in papers that combine the two methods include surveys of users describing recent trips, which are still self-reported and subject to errors in memory or perception and to exaggeration. Revealed preference data on its own also comes with many limitations, and studies that use automated data are still rare. Only one of these studies included choice models based on automated rather than self-reported data. For revealed preference data, the shortcomings include the fact that only the 'most preferred' option is reported, there may be correlation among the different variables, there may be a lack of variation in the data, and important factors may be excluded. In the case of this dissertation, variables representing traffic conditions may be correlated with toll amounts and time savings, for example. Peak-hour toll rates could potentially be consistent, reducing the variation in

the data. Trip attributes, such as trip purpose and desired arrival time, are not captured by the automated data and may affect modeling results. These issues are discussed later in the limitations section. The next chapter will discuss the data used in this dissertation. It will give an overview of the various data streams that are ultimately used to assess trip, corridor, and household characteristics, as well as lane choice decisions.

CHAPTER 3

DATA SOURCES

The data supporting this dissertation come from two main sources: Express Lane use and performance data provided by the Georgia State Road and Tollway Authority (SRTA), and socioeconomic data provided by the marketing firm Epsilon. This chapter will begin by outlining the SRTA data streams and providing descriptive overviews of the SRTA data that have been collected. This includes a thorough investigation of the lane use data and the patterns of behavior that they reveal. The next section will describe the Epsilon marketing data. That section will describe the coverage of the demographic variables provided by Epsilon and will also provide distributions of a subset of those variables. The final section examines the correlation among the Epsilon demographic variables.

SRTA Express Lane Data

The I-85 Express Lane data that SRTA provides consist of ten different streams that are delivered automatically to Georgia Tech's servers. The streams come from ETCC, the contractor that works with SRTA to operate the Express Lanes. The data arrive as XML files, with data frequencies varying from every twenty seconds to every seven days. This section gives an overview of the elements, frequency, and significance of the primary data streams that will be used in this research. Descriptions of the remaining data streams can be found in the Appendix.

As discussed earlier in this dissertation, the physical infrastructure that makes up the HOT system includes thirty-five Express Lane gantries positioned approximately half

a mile apart. One of these gantries sits above SR-316 to identify the users of that portion of the facility. In addition to the Express Lane gantries, Radio Frequency Identification (RFID) scanners detect vehicles with Peach Pass transponders in the General Purpose lanes. Seven of these operate in the northbound direction and six in the southbound direction. These segments are separated by weave zones that are between 0.35 miles and 0.82 miles long. Each travel direction also has five rate signs that display the toll amount required for the next segment of the facility and the amount required to travel to the end of the facility. These signs are placed before each of the weave zones that allow drivers to enter or exit the HOT lanes. The Express Lane gantries and AVI scanners provide the detection data described in the following sections.

Primary Data Streams:

Vehicle RFID Tag Read Data: This stream contains all vehicle detections for all lanes. The automatic vehicle identifiers provide information about the transponder, the vehicle, and the lane the vehicle was traveling in. In addition to the TransponderID and the LaneID, the table provides a timestamp column to identify which lane a specific vehicle was in at a given time. These data have many uses, including travel time and reliability calculations and weaving studies.

Frequency: Delivered daily.

Data Fields:

Table 2: Vehicle Read Data Elements

Name	Description
TransactionID	Unique Transaction Identifier
TransponderID	Unique Transponder Identifier
PlateNumber	License plate of vehicle
PlateState	Registration state of vehicle
LaneID	Unique Lane Identifier
TransactionDateTime	Timestamp of vehicle detection

The vehicle stream delivers an enormous amount of data. In 2012 alone, there were over 78 million detections in the HOT lanes (roughly 2.3 per mile) and over 44 million detections in the General Purpose lanes (roughly 0.4-0.5 per mile). Table 3 below provides an overview of the vehicle data stream for 2011, 2012, 2013, and 2014. The "Total HOT/GP Detections" rows report the raw numbers of detections from each lane type provided over the year. "Average Unique HOT/GP Users per Month" are derived from the monthly counts of distinct transponders detected in each lane type. The values indicate, as may be expected, that more Peach Pass transponders are detected in the General Purpose lanes versus the HOT lanes.

Table 3: Vehicle Stream Summary

	2011 (Oct-Dec)*	2012 (Jan-Dec)	2013 (Jan-Dec)	2014 (Jan-Dec)
Total HOT Detections	13,220,332	78,340,186	94,974,194	108,718,150
Total GP Detections	0	44,368,481	62,159,534	63,614,769
Average Unique HOT Users per Month	29,315	48,476	63,328	73,337
Average Unique GP Users per Month	0	142,259	176,170	180,515

*: Prior to January 6, 2012, the General Purpose vehicle detectors were not operating.

Figure 3 illustrates the number of unique users per month in each lane type from the opening of the facility in October, 2011 through December of 2014. The jump in GP lane users in January, 2012 is the result of the GP lane vehicle detectors coming online;

prior to that month, only the HOT lane detectors were reporting data. Both lines illustrate the gradual increase in Peach Passes detected on the corridor since the start of operations.

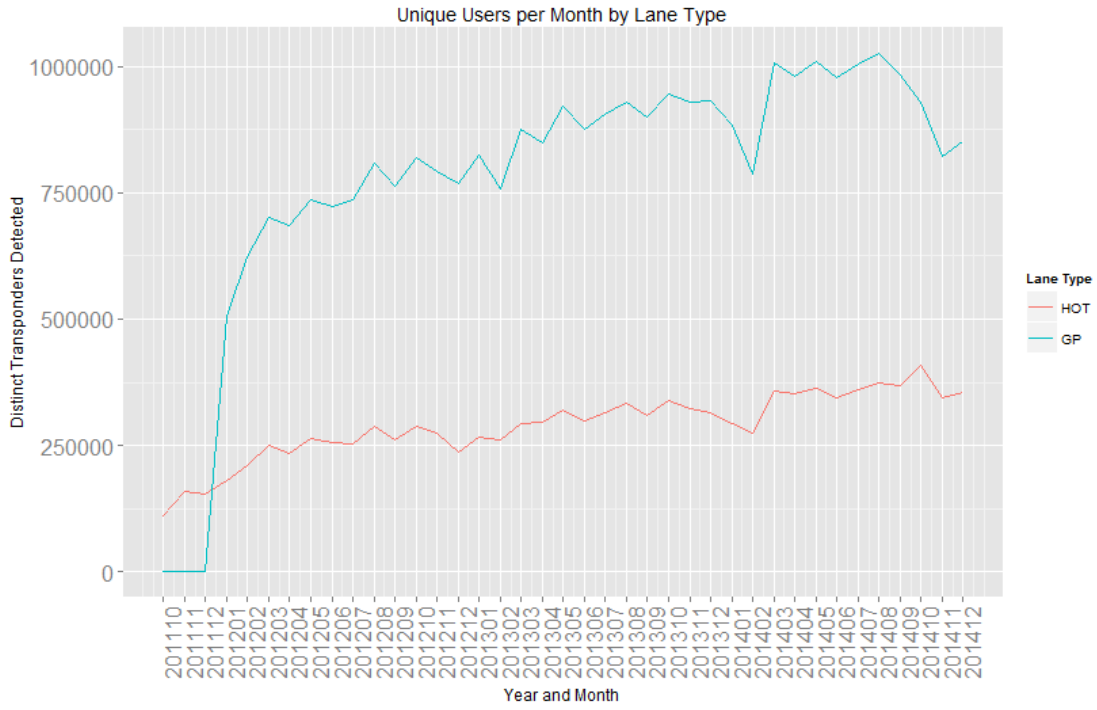


Figure 3: Users per Month by Lane Type

An important behavioral feature for the analysis methods that are implemented later in this dissertation is the fluidity of lane type choices; that is, users do not remain in only the HOT or GP lanes. Table 4 below shows the numbers of transponders that are detected in each lane type in each month of 2012. The significant result here is that over 30,000 vehicles use both the HOT and General Purpose lanes each month. These "hybrid" users make up between 28.1% and 33.4% of the unique corridor users each month. Similar tables for 2013 and the first three months of 2014 can be found in the Appendix.

Table 4: Unique Corridor Users by Lane Type for 2012

Month	GP-Only Users	HT-Only Users	GP and HT "Hybrid" Users	Total Users	Percentage "Hybrid" Users
201201	76,791	2,024	30,834	109,649	28.1%
201202	84,518	2,453	35,099	122,070	28.8%
201203	92,368	3,170	40,419	135,957	29.7%
201204	93,157	3,268	40,902	137,327	29.8%
201205	95,193	3,719	45,718	144,630	31.6%
201206	99,396	3,821	46,425	149,642	31.0%
201207	101,342	3,831	47,525	152,698	31.1%
201208	100,707	4,017	50,546	155,270	32.6%
201209	100,402	3,800	49,925	154,127	32.4%
201210	104,294	2,746	50,400	157,440	32.0%
201211	104,550	3,057	54,081	161,688	33.4%
201212	111,515	2,927	50,991	165,433	30.8%

Trip Data: The trip data stream contains information on all Express Lane trips that occurred within the past day. The records provide the section in which the trip took place, the unique RFID identifier of the transponder detected, the toll amount paid, the license plate number, entry and exit times for the HOT lane, and an indicator of whether the trip occurred in toll-paying or carpool mode. With these data elements, researchers can determine when a specific vehicle entered the HOT lane from the General Purpose lanes, the length of the trip, and what price they paid both overall and per mile.

Frequency: Delivered daily.

Data Fields:

Table 5: Trip Data

Name	Description
TripID	Unique Trip Identifier
SectionID	Unique Section Identifier
TollAmount	Amount paid
TollMode	TOLL or NON-TOLL
TransponderID	Unique Transponder Identifier
PlateNumber	License plate of the vehicle
PlateState	Registration state of the vehicle
TripEntryTime	Timestamp of start of the trip
TripExitTime	Timestamp of end of the trip
DWLViolationFlag	'Y' if trip involved a double white line violation

Table 6 shows a summary of many of the data elements of the Trip stream for the two months of operations in 2011 and all of 2012, 2013, and 2014. "Paid trips" refers to trips taken in toll mode (as opposed to non-toll mode) with toll amounts greater than zero. A trip may be registered with a toll Amount of $0 if the operating agency overrode the system, likely due to a "breakdown" of toll lane conditions. The vast majority of the trips, over 90%, are taken in 'Toll' mode. In the trip data stream, 'Non-Toll' mode is used to describe carpool trips that do not get charged as well as trips by emergency vehicles, alternative fuel vehicles, and motorcycles.

Table 6: Trip Stream Summary

	2011 (Nov-Dec)*,**	2012*	2013	2014
Average Toll Amount Paid (All Trips)	$0.94	$1.00	$1.21	$1.44
Average Toll Amount Paid (All Paid Trips)	$1.14	$1.11	$1.47	$1.71
Average Toll Amount Paid (Peak Hour and Direction Paid Trips)	$1.41	$2.08	$2.93	$3.39
Toll Mode Trips Percentage	91.69%	93.66%	93.72%	93.09%
Average Total Trips per Month	201,904	338,343	411,390	465,137
Total Trip Records	403,808	4,060,112	4,936,680	5,581,643

*: Prior to 01/29/2012, five trip sections were not included in the data. Those sections were PHS-PHS, PHS-ITS, PHS-JCS, PHS-285S, and OPS-OPS.

**: Trip stream data were not delivered for October of 2011

Table 7 shows average southbound toll amounts broken down by day of the week and morning peak hour for all of 2012. Table 8 shows the corresponding average tolls for the northbound afternoon peak. Both tables reflect paid toll trips, as discussed above. Average southbound morning peak tolls are higher than their northbound afternoon peak counterparts with the exception of the northbound 6:00pm hour and the 5:00pm hour on Fridays. Note that these averages include all paid toll trips; they do not control for the impact of trip length on the toll amount charged. Tables for calendar years 2013 and 2014 can be found in Appendix A.

Table 7: 2012 SB AM Peak Average Tolls by Day of Week and Hour – Paid Trips

n = 1,146,473 trips	6:00 AM	7:00 AM	8:00 AM	9:00 AM	AM Peak
Monday	$2.05	$2.90	$2.19	$1.03	$2.23
Tuesday	$2.09	$2.97	$2.36	$1.17	$2.31
Wednesday	$2.09	$3.01	$2.25	$0.96	$2.27
Thursday	$2.06	$2.92	$2.23	$1.05	$2.24
Friday	$1.51	$1.90	$1.32	$0.72	$1.48
All	$1.98	$2.77	$2.10	$1.00	$2.13

Table 8: 2012 NB PM Peak Tolls by Day of Week and Hour – Paid Trips

n = 1,095,483 trips	3:00 PM	4:00 PM	5:00 PM	6:00 PM	PM Peak
Monday	$0.91	$1.12	$1.34	$1.13	$1.16
Tuesday	$0.91	$1.20	$1.41	$1.17	$1.21
Wednesday	$0.91	$1.16	$1.33	$1.13	$1.16
Thursday	$0.93	$1.22	$1.45	$1.29	$1.26
Friday	$0.98	$1.26	$1.41	$1.20	$1.23
All	$0.93	$1.19	$1.39	$1.19	$1.21

Figure 4 illustrates total Express Lane trip counts by month, including both paid and unpaid (carpool mode) trips. The largest contributor to the variation in total trip counts each month is the number of toll-mode trips; non-toll trips increase very gradually from the start of operations through the end of 2014. The chart reveals a gradual increase in the total number of HOT lane trips per month, punctuated by occasional steep declines and increases.

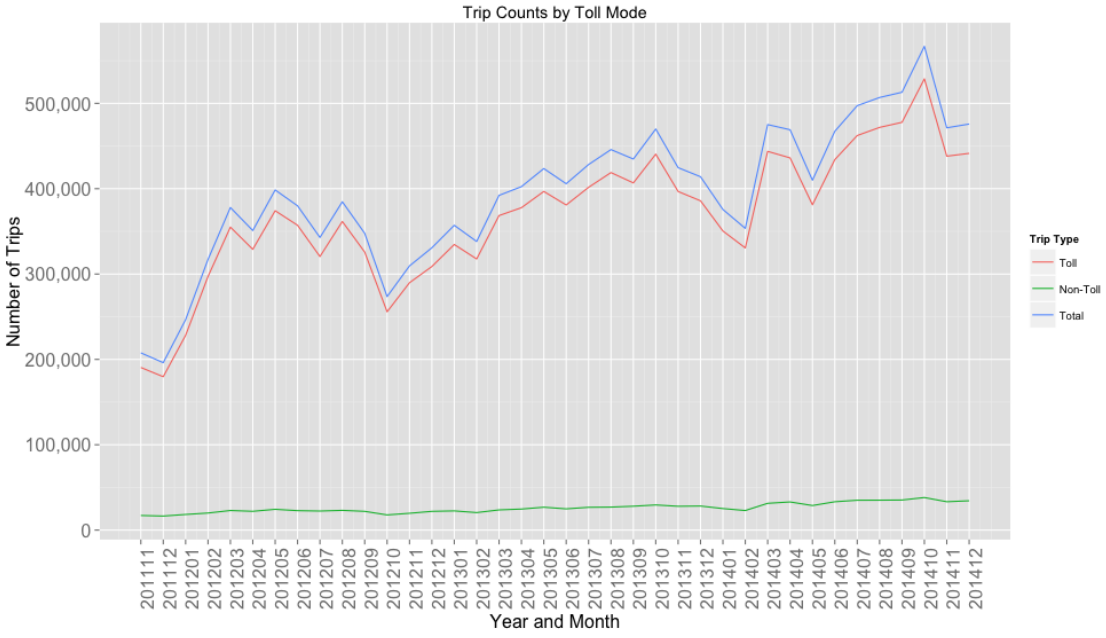


Figure 4: Trip Counts by Month

Toll Trip Overview

Figure 5 illustrates the consistency of toll-mode trip taking: the ratio of toll mode to non-toll mode is virtually unchanged for the duration of the study time period. Figure 6, which illustrates the total toll amounts paid per month as reported by the Trip stream, also indicates gradual growth across the entire timeframe. This growth is an effect of both the increasing number of trips (as seen in Figure 4) and the increase over time of the maximum possible toll rate for a given trip. This toll rate increase will be illustrated later in Figure 13.

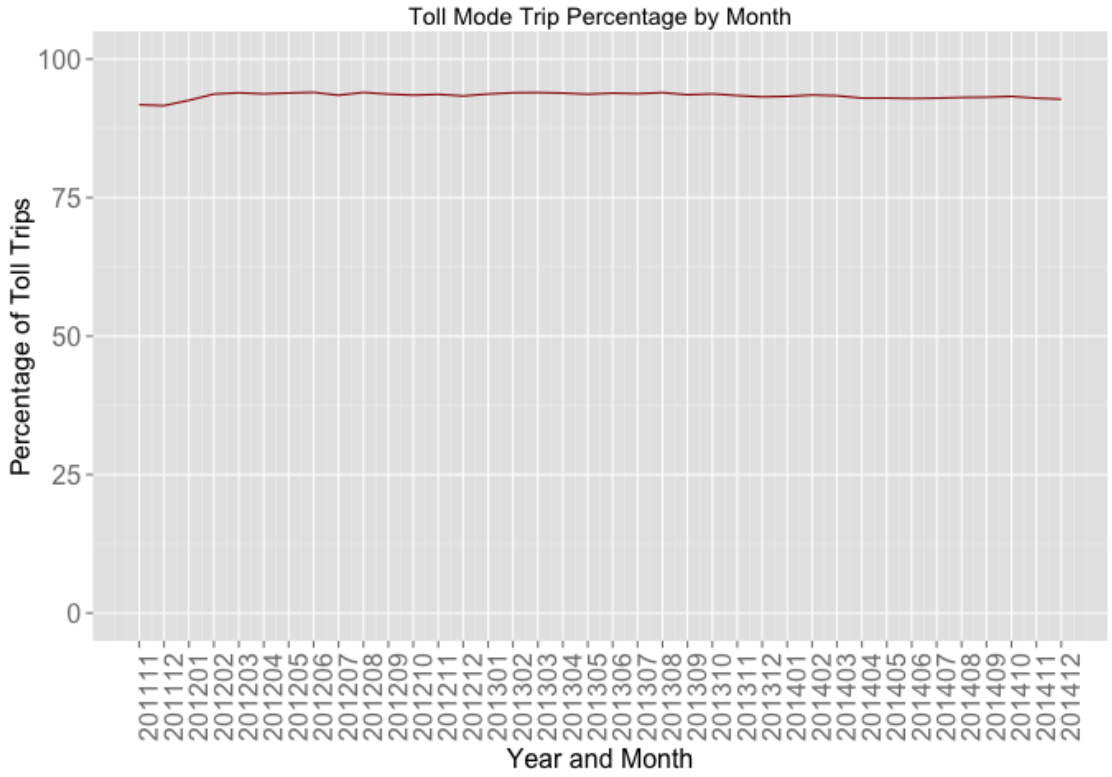


Figure 5: Toll Mode Trip Percentages by Month

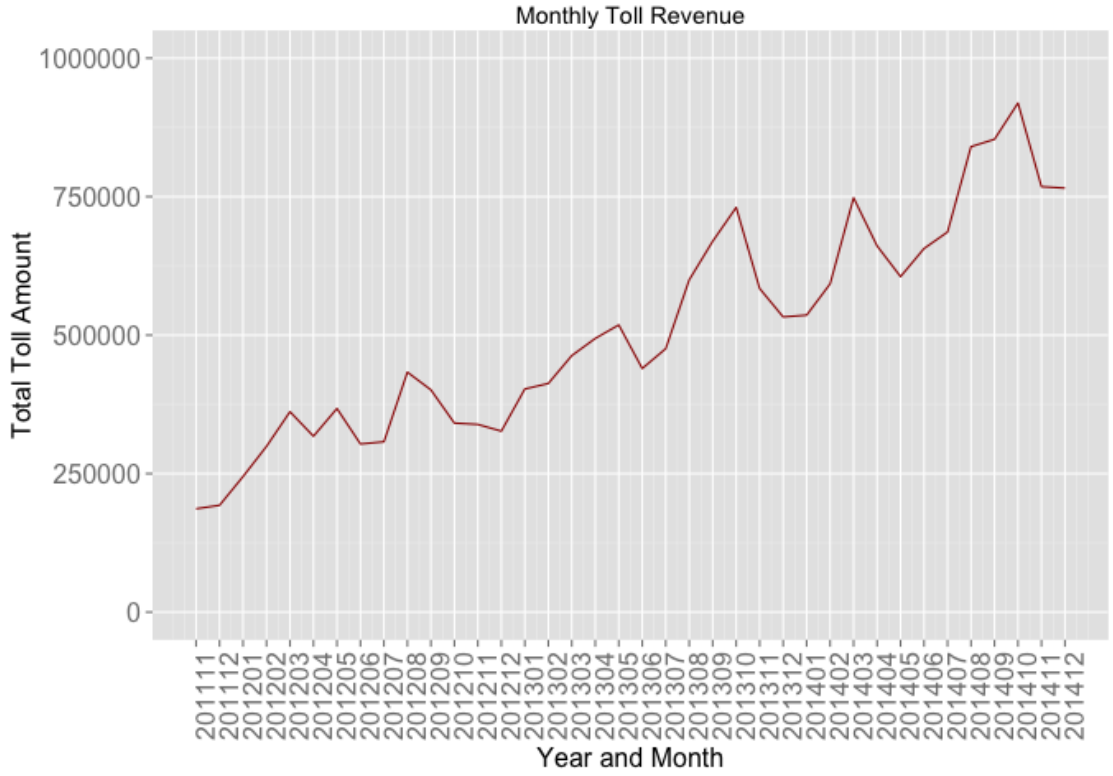


Figure 6: Monthly Toll Revenue Since Inception

Figure 7, Figure 8, and Figure 9 below show the distributions of tolls paid during all hours and during the AM and PM peak hours in 2012. The distributions illustrate the variation in toll amount paid, even within the peak periods. Here again the trips under examination are 'paid' trips, which are in toll mode and have toll amounts greater than zero. Southbound AM peak trips exhibit a greater variety in potential toll amounts; this is reflected in the shape of the distributions and the higher median toll for southbound trips. Similar charts for 2013 and 2014 can be found in the Appendix.

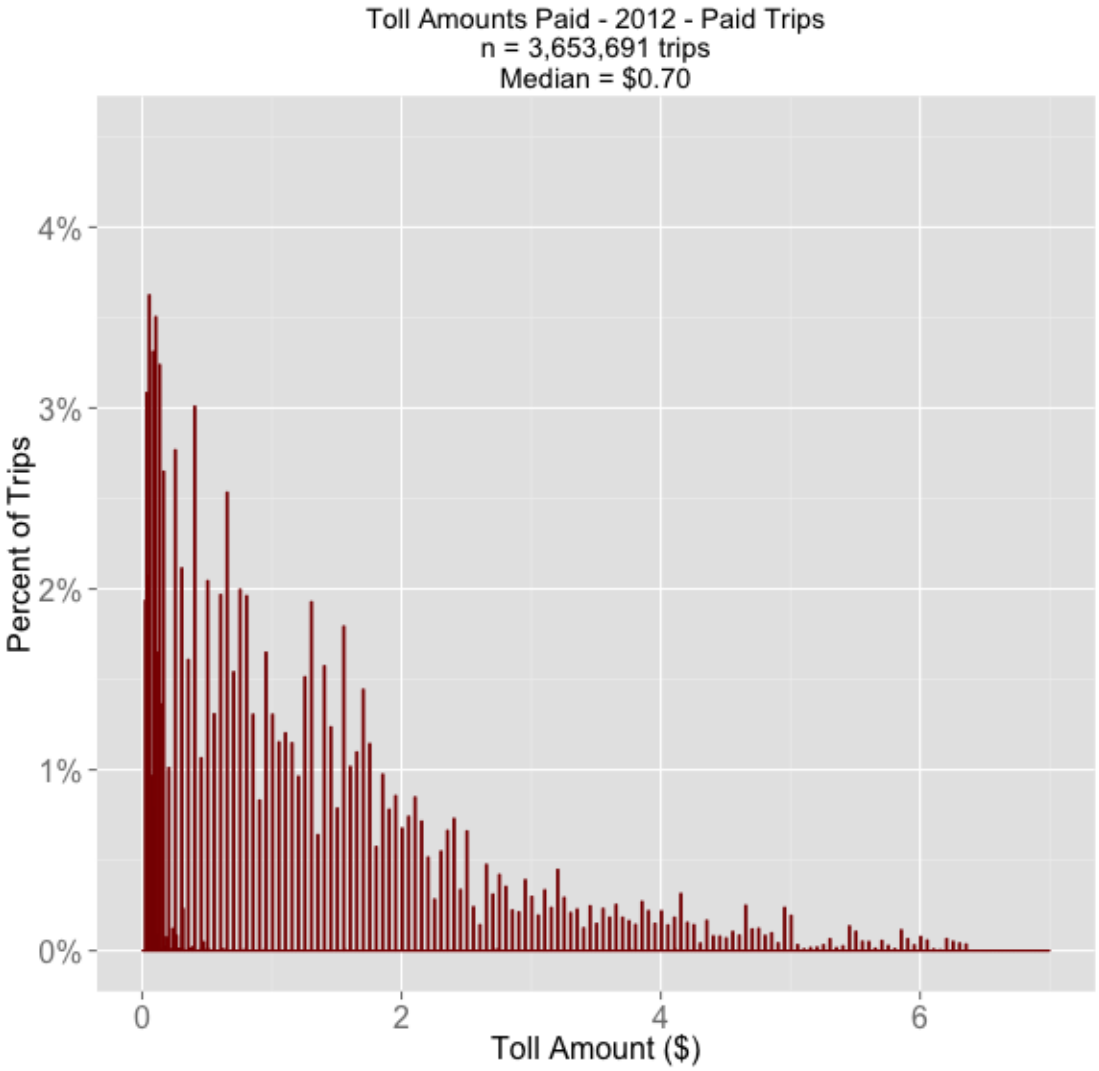


Figure 7: 2012 Distribution of Paid Tolls

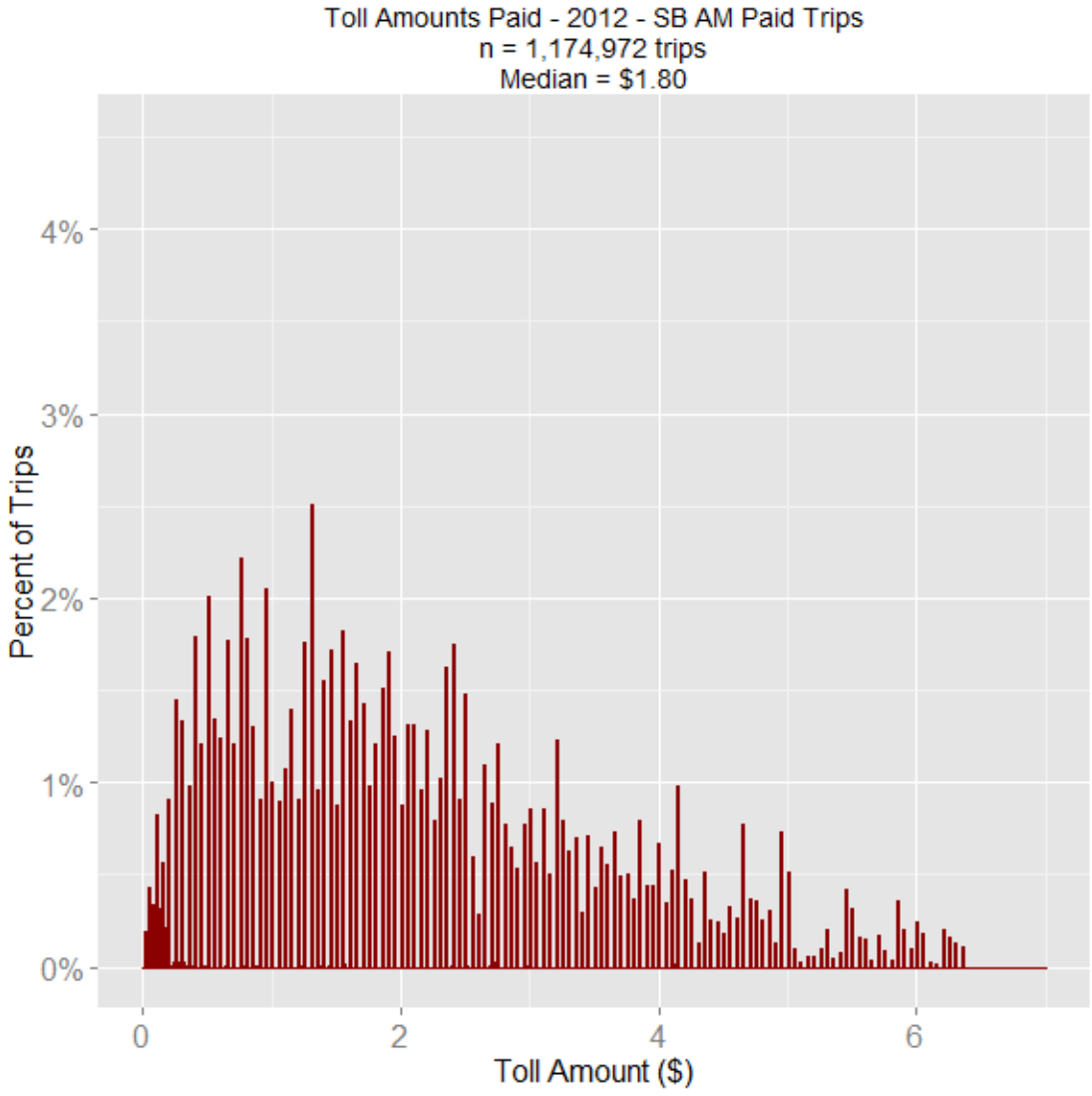


Figure 8: 2012 Distribution of Paid Tolls, Southbound AM Peak

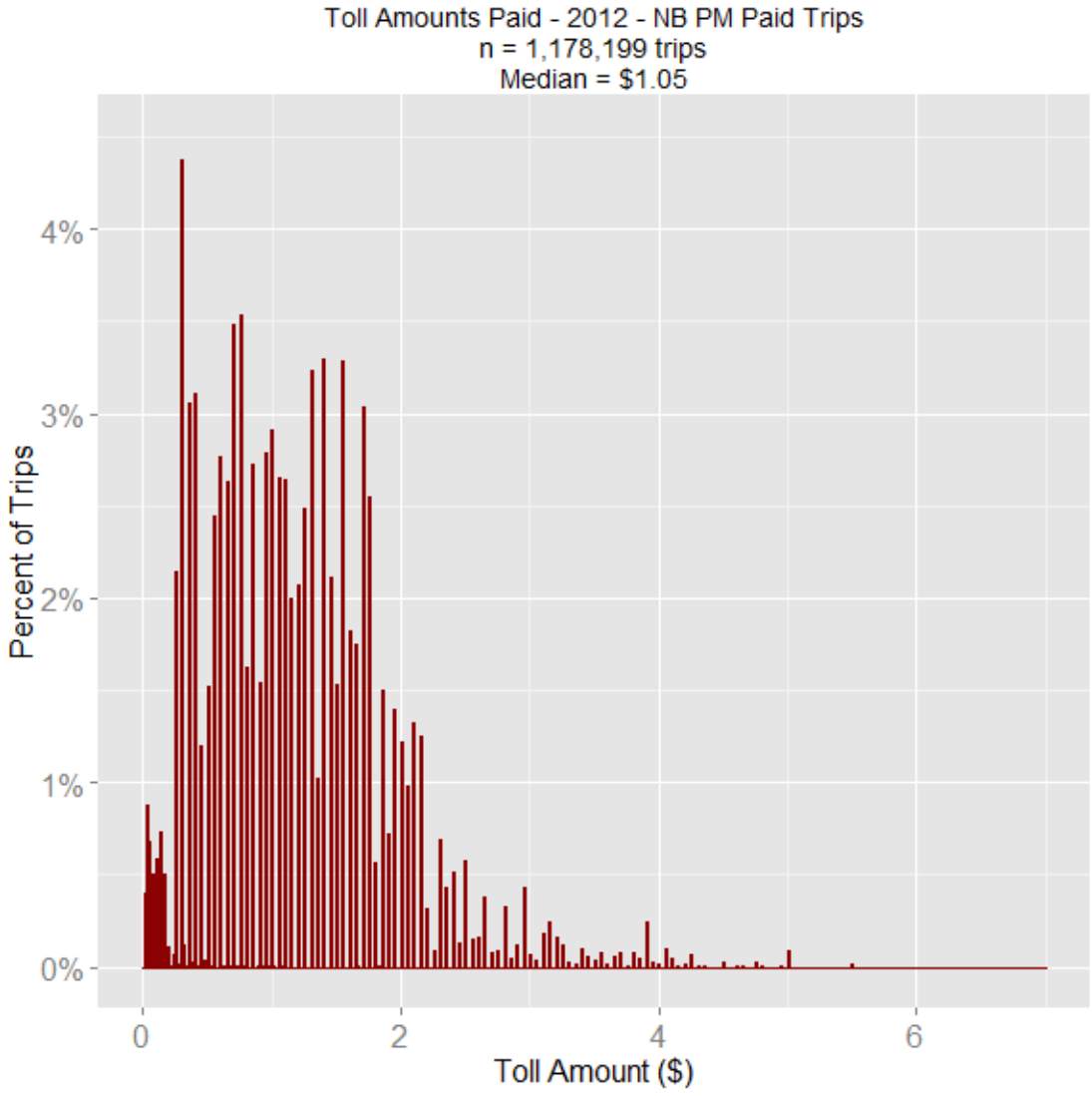


Figure 9: 2012 Distribution of Paid Tolls, Northbound PM Peak

Figure 10 and Figure 11 illustrate the toll amount distributions only for paid trips that traverse the entire corridor: Southbound from Old Peachtree Road to I-285 in the morning peak (6:00 AM to 10:00 AM), and northbound from I-285 to Old Peachtree Road in the afternoon peak (3:00 PM to 7:00 PM). Again, the southbound morning peak trips exhibit more variation and have a higher maximum toll amount; the northbound afternoon trips are more tightly clustered around the median and do not exceed $6. Similar plots for 2013 and 2014 can be found in the Appendix A.

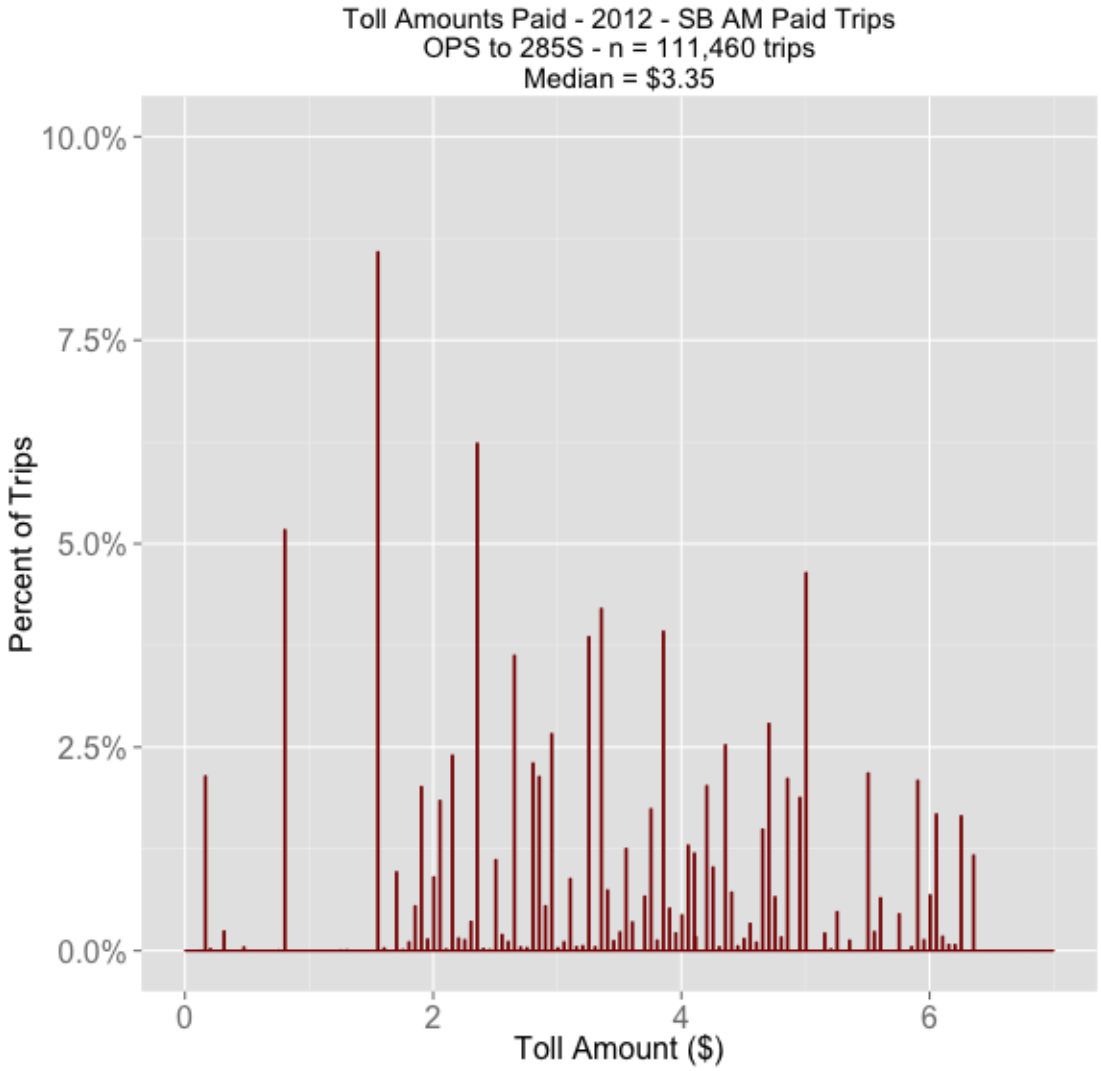


Figure 10: 2012 Distribution of Paid Tolls, Southbound AM Peak - Section 23

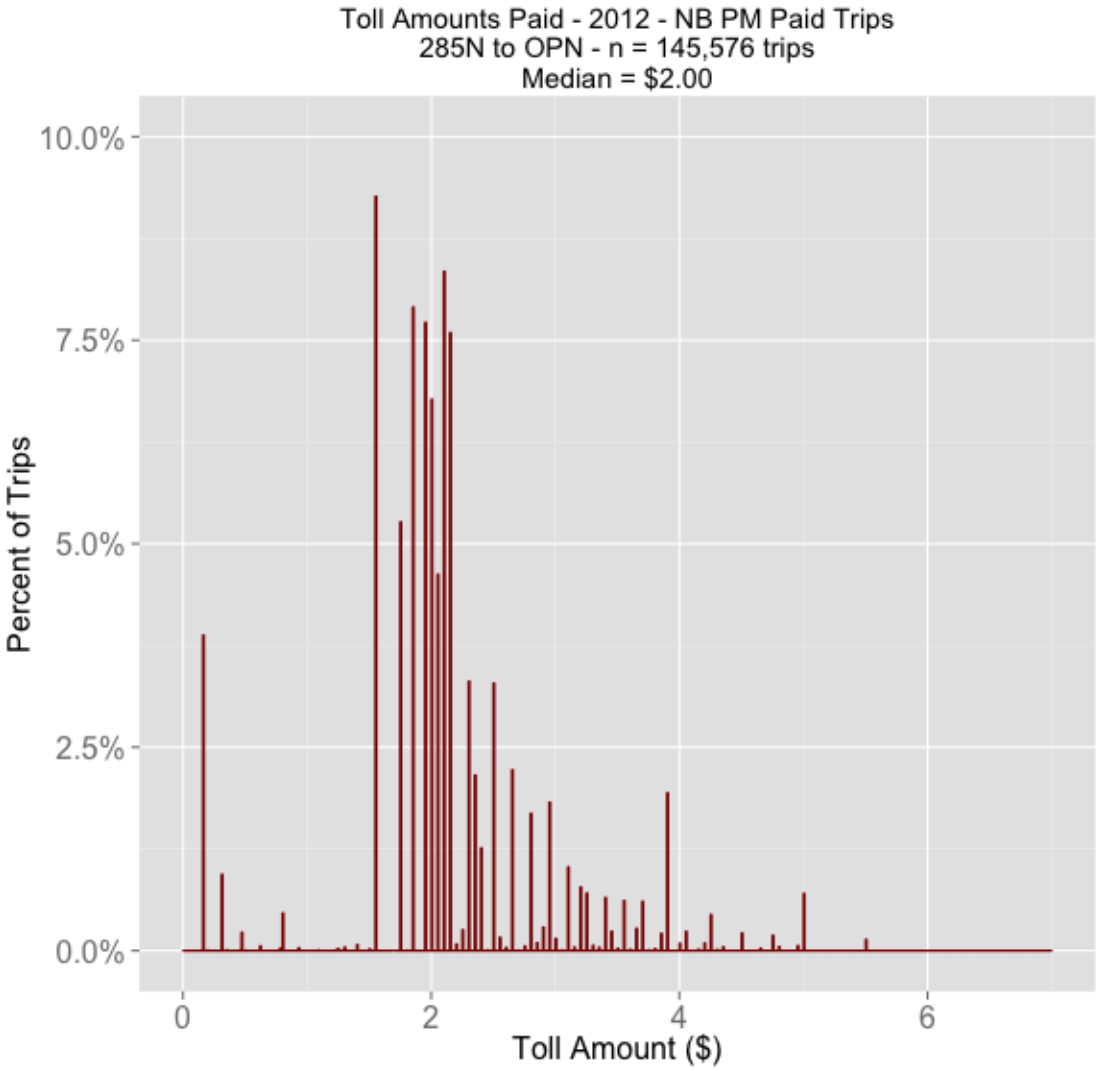


Figure 11: 2012 Distribution of Paid Tolls, Northbound PM Peak - Section 5

Figure 12 shows the average toll amount charged per month for the peak-period, corridor-length, 'paid' trips discussed above. Both the southbound and northbound lines slowly trend upwards, with the northbound toll amounts increasing rapidly at the end of 2013. There are multiple potential reasons for a constantly increasing toll rate, including higher demand on the corridor and decreasing sensitivity to toll amounts.

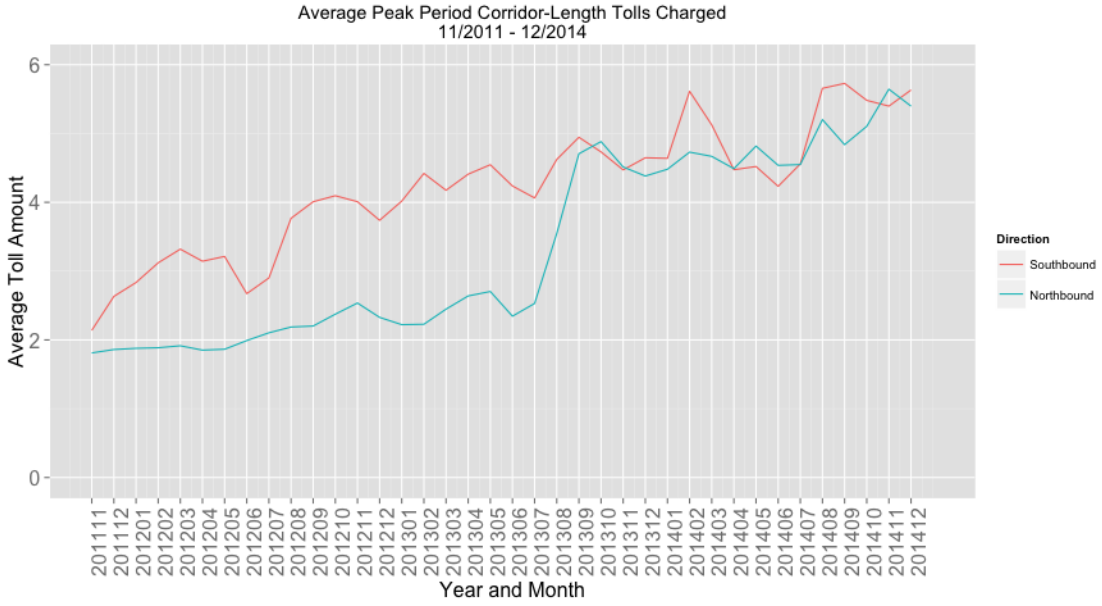


Figure 12: Average Peak Tolls Charged by Month

One complication with the amounts charged by SRTA and reported in the trip stream is the slowly-increasing upper limit imposed on the tolls. Political considerations resulted in SRTA capping the maximum allowable toll, with that cap increasing over time. Evidence of this gradual increase in the maximum toll can be seen here in Figure 13. The implications of this changing toll cap are described in later sections of this dissertation.

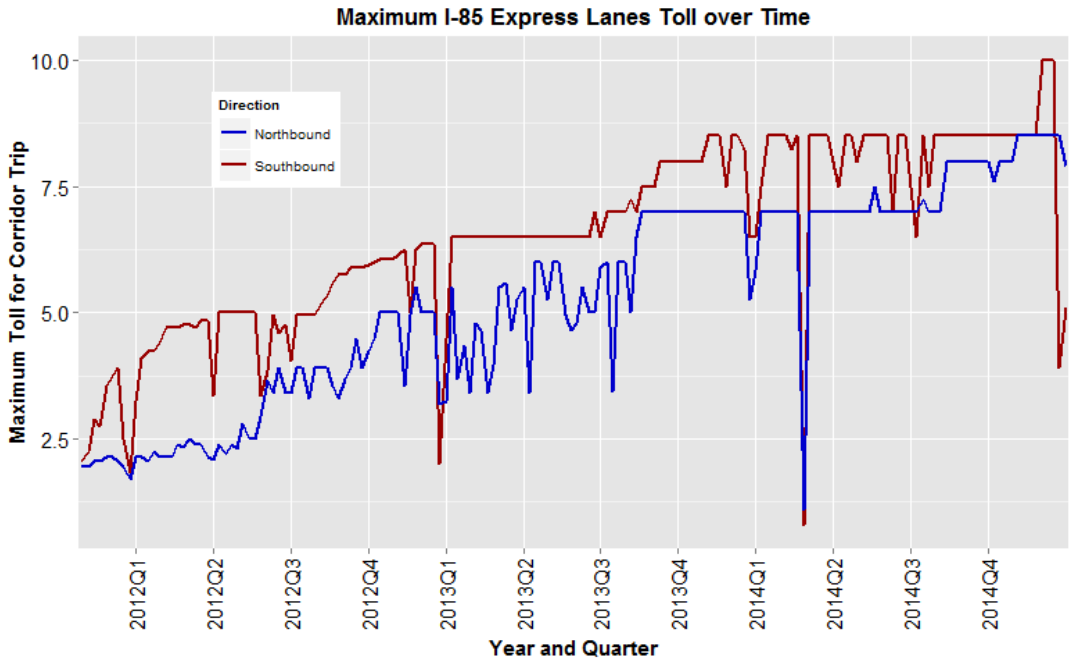


Figure 13: Maximum Toll Charged per Week

Figure 14, Figure 15, Figure 16, and Figure 17 take this issue into account by presenting distributions of the toll amount paid as a fraction of the maximum toll for a given time frame. Figure 14 and Figure 15 illustrate the distribution of tolls paid as a fraction of the weekly maximum, while Figure 16 and Figure 17 look at the daily maximum. These are peak period, peak direction trips that traverse the entire corridor, from Old Peachtree to 285 and vice versa. In both the weekly and daily charts, the pluralities of trips occur at the maximum toll rate. This is even more apparent at the daily level: over 30% of the southbound AM trips occur at the maximum toll rate for the day, while for northbound PM trips that figure is over 20%.

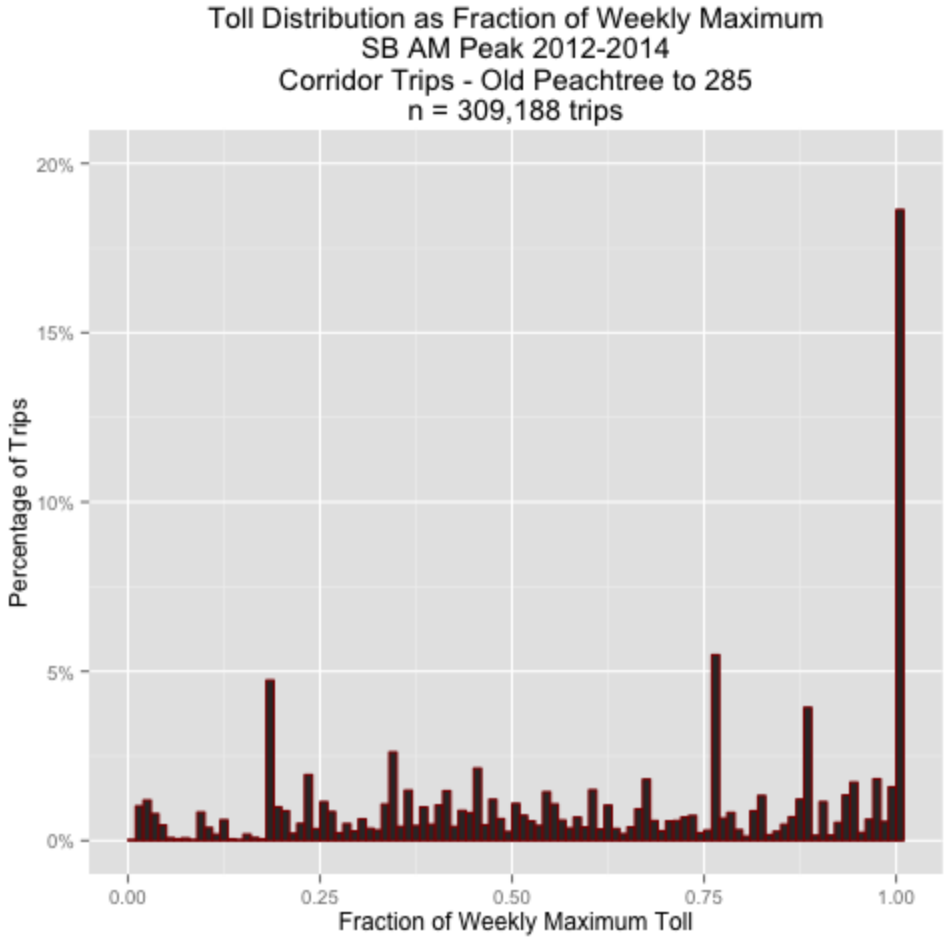


Figure 14: Toll Distribution as Fraction of Weekly Maximum, SB 2012

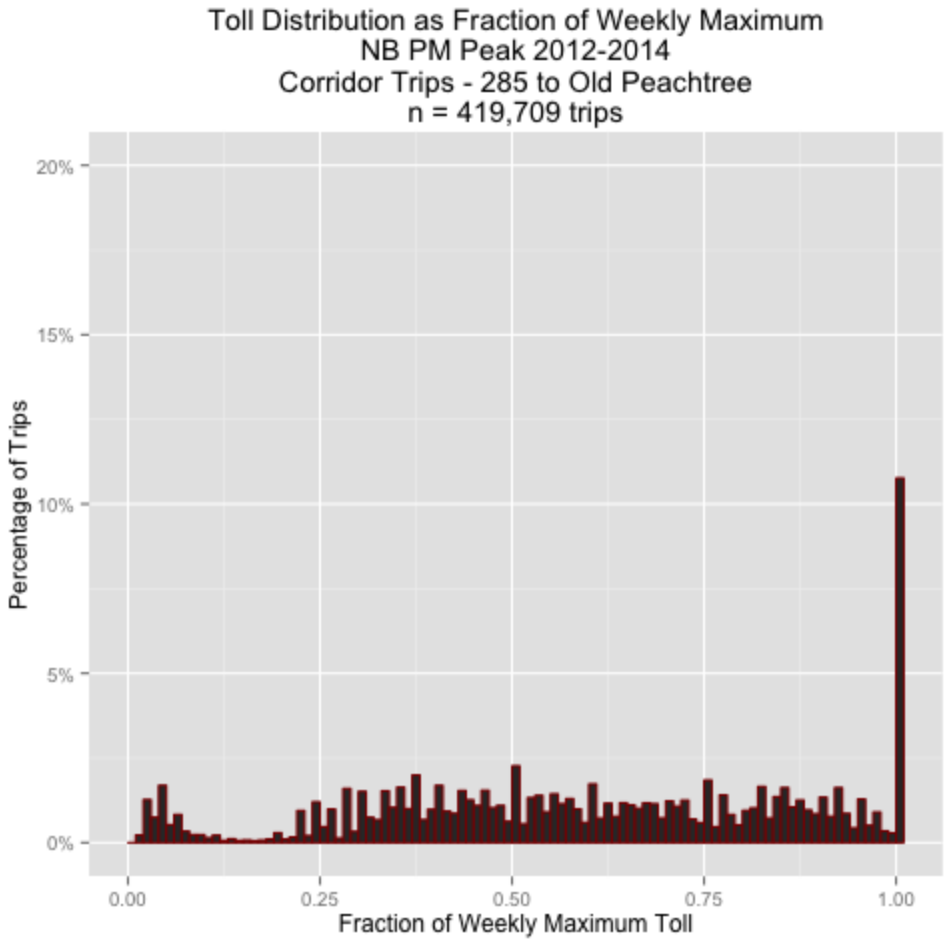


Figure 15: Toll Distribution as Fraction of Weekly Maximum, NB 2012

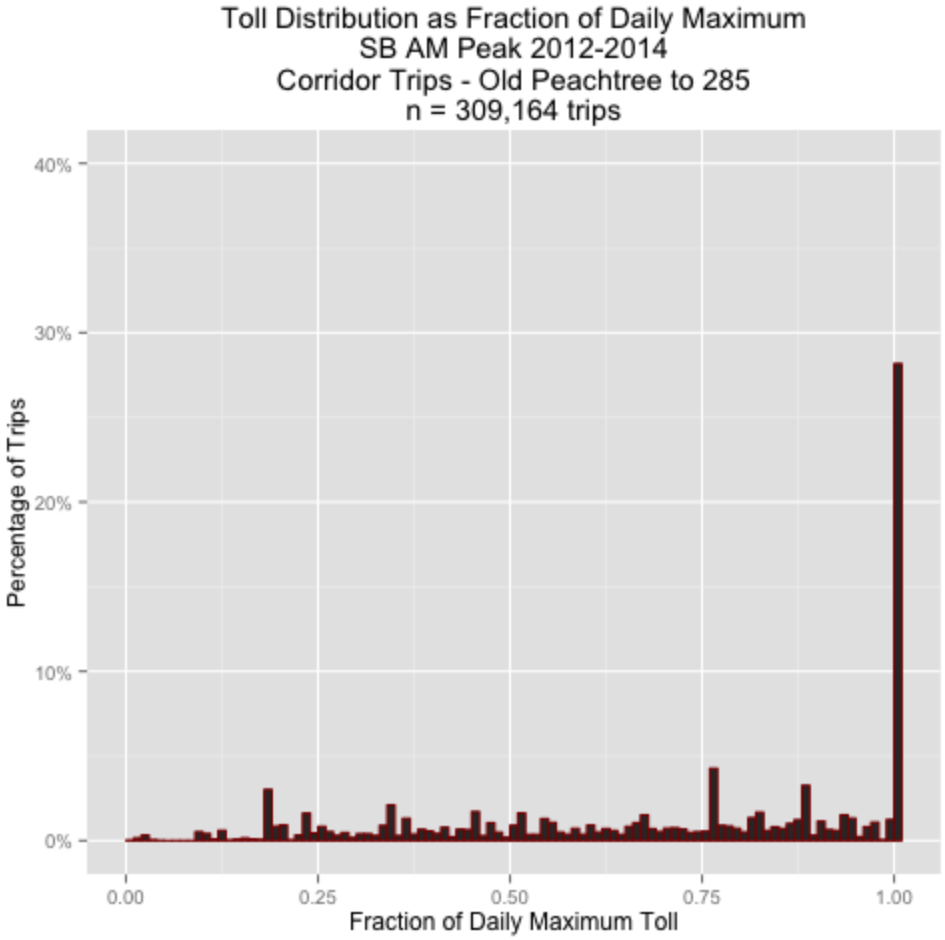


Figure 16: Toll Distribution as Fraction of Daily Maximum, SB 2012

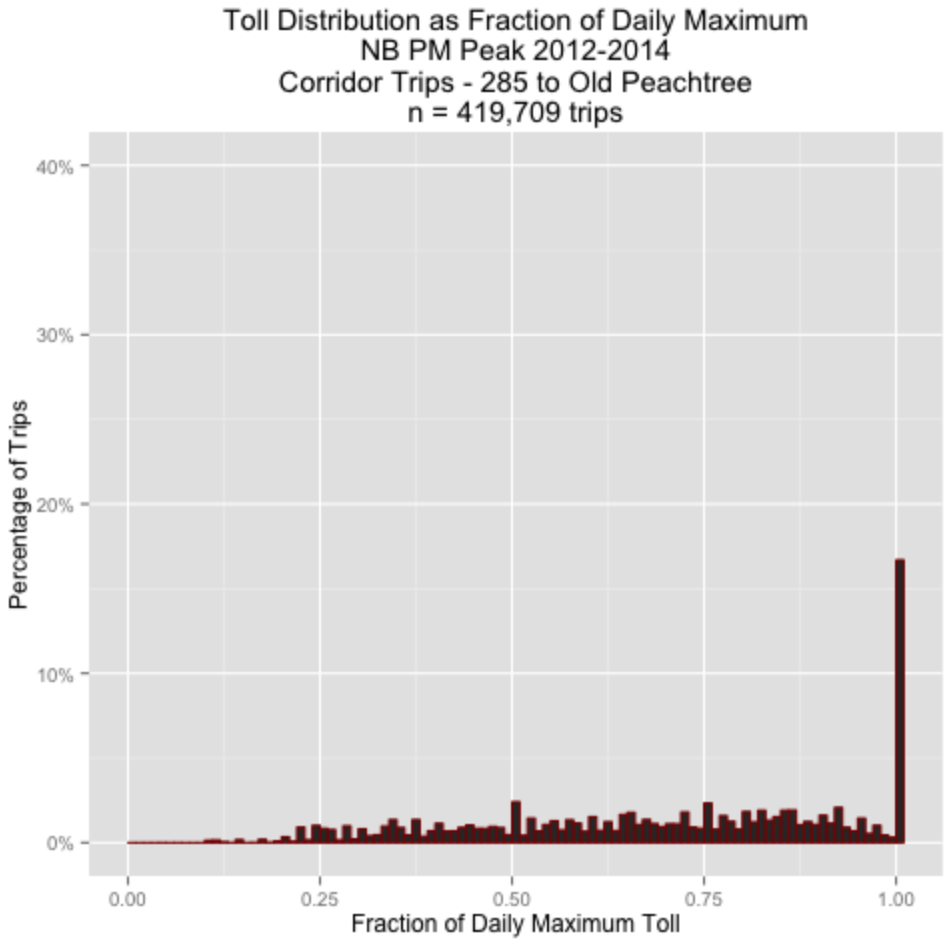


Figure 17: Toll Distribution as Fraction of Daily Maximum, NB 2012

These charts illustrate a notable feature of Express Lane trips: a large proportion (roughly 10-30%) of these trips occur at the maximum toll rate for a given day. This indicates that the toll is insufficiently high to meet the demand management goals of the facility. Under these conditions, congestion in the toll lanes becomes more likely as users are not sufficiently discouraged from purchasing trips.

Transponder Trip Distributions

Figure 18 shows the distribution of trips by Peach Pass transponders in 2012 through 2014. Roughly 20% of the transponders registered only a single trip in the entire year. This distribution is based on transponders that appeared in the Trip data, and so it does not include those that did not use the Express Lanes in 2012. This first figure includes all trip records, including those in both Toll and Non-Toll (carpool) modes. In addition, it includes trips by both personal and corporate accounts.

Figure 19 looks at paid trips only; that is, trips that occur in Toll mode and were charged an amount greater than $0. The distributions are very similar in shape, with a slightly higher peak at one trip and a decrease in the median of one for the Paid Trips distribution. Figure 20 illustrates the trip distribution of Non-Toll trips which were charged $0. Again, transponders with only one trip make up the plurality of the data set, but here the discrepancy between one and two trips is not as great.

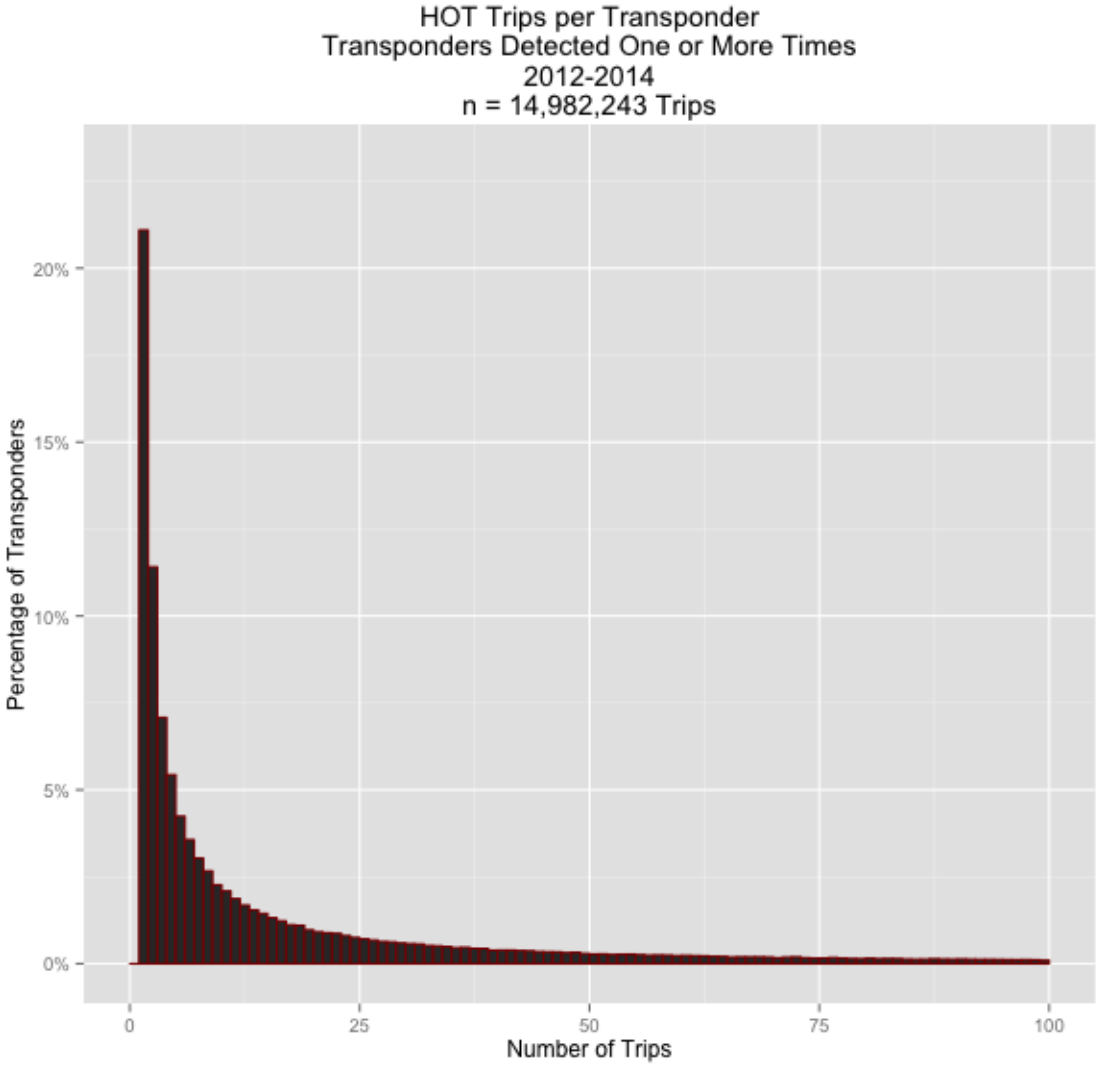


Figure 18: HOT Trips per Transponder for 2012-2014

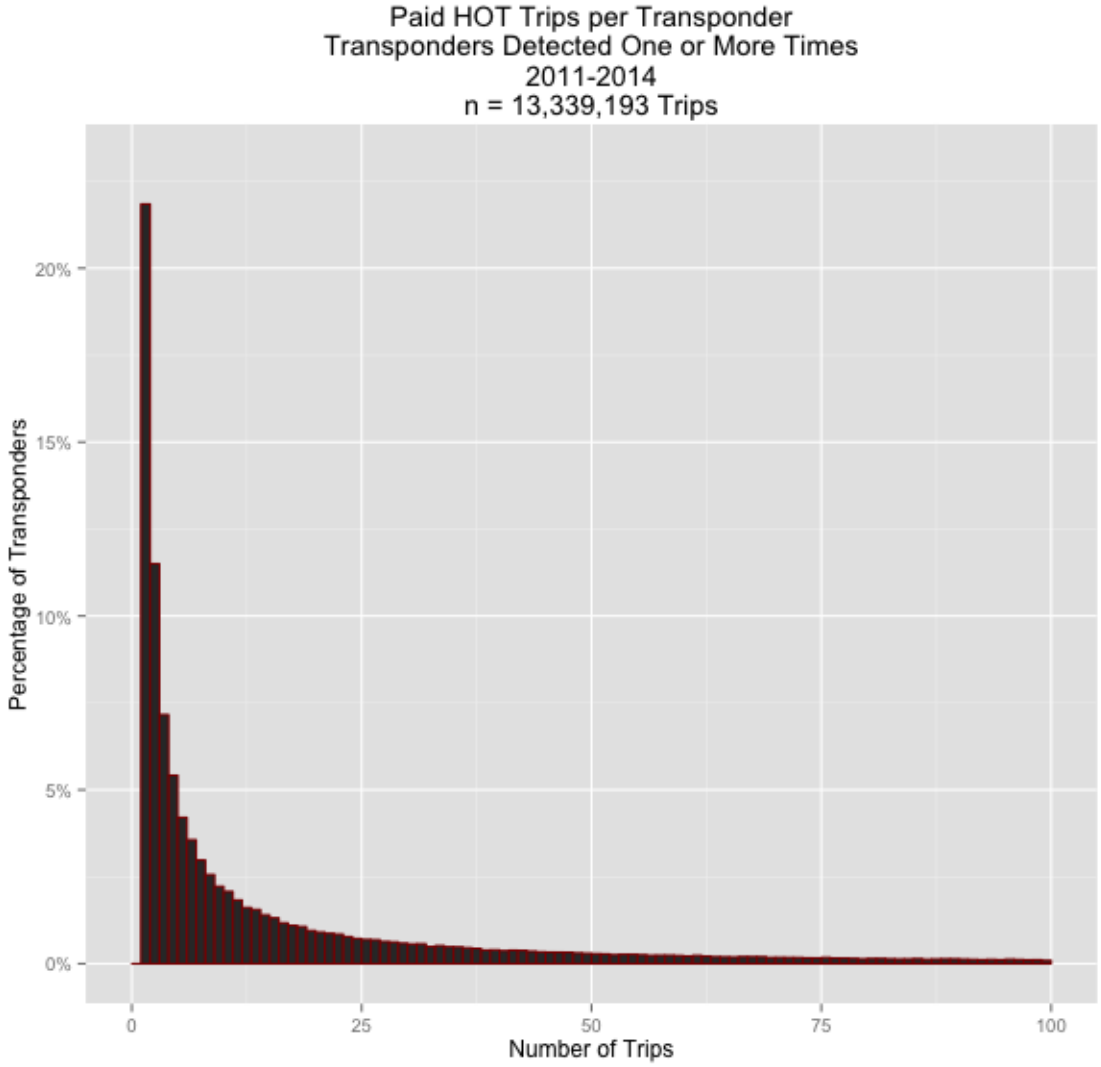


Figure 19: Paid HOT Trips per Transponder for 2012-2014

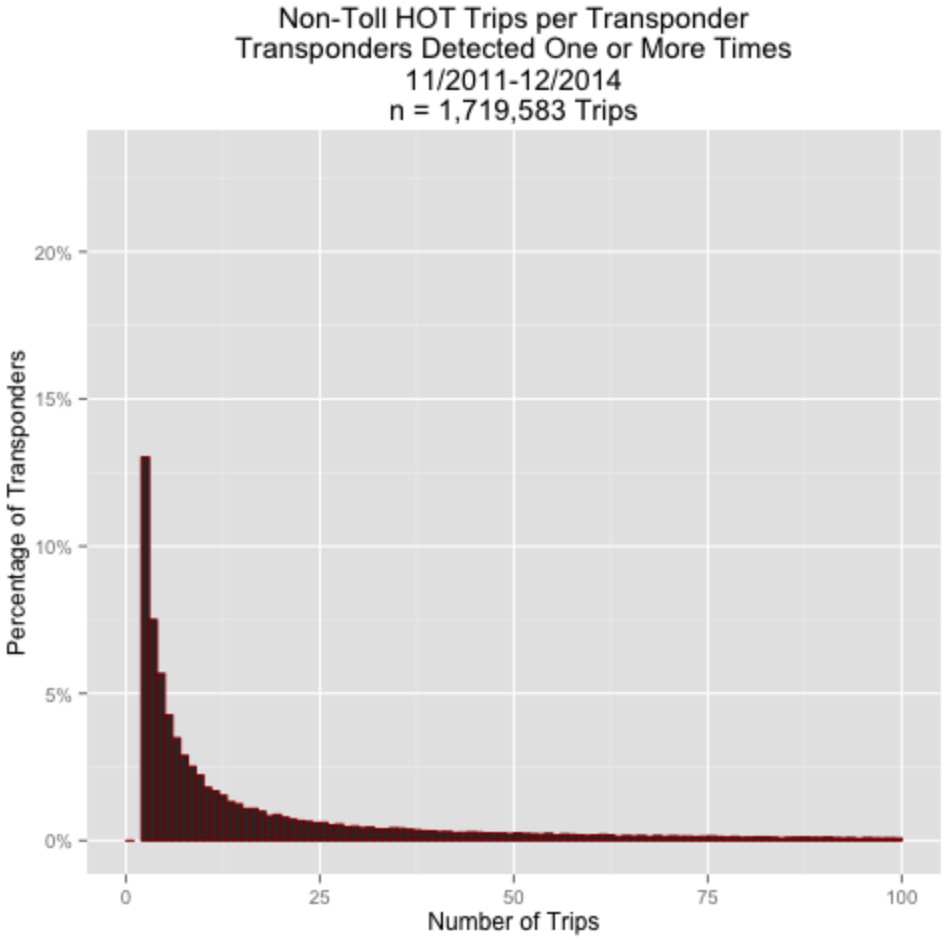


Figure 20: Non-Toll HOT Trips per Transponder for 11/2011-12/2014

Figure 21 below presents the distribution in a different form, in which the transponders have been ranked by the total number of trips in 2012. This chart illustrates that the 20% of transponders with the most trips undertook more than 80% of the total trips. Again, this sample is limited to transponders that had at least one HOT lane trip in 2012 and thus appeared in the trip data stream. As discussed above, the trips counted here include both toll and non-toll (carpool) trips, as well as trips by both personal and corporate accounts.

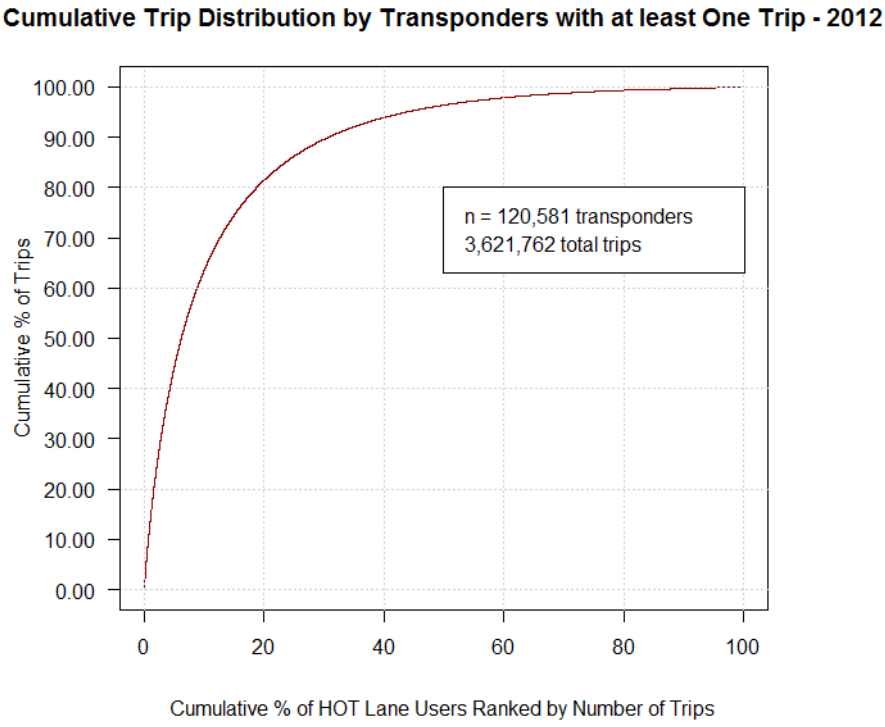


Figure 21: Cumulative Trip Distribution for 2012

Figure 22 shows the distribution of paid trips by transponder, excluding Non-Toll carpool trips and those with toll amounts of zero. Finally, Figure 23 illustrates the distribution of Non-Toll trips per transponder. All three figures share a very similar shape despite differences in total numbers of transponders and trip counts; in all three cases, the top 20% of transponders took roughly 80% of the trips. Similar distributions for calendar years 2013 and 2014 can be found in the Appendix.

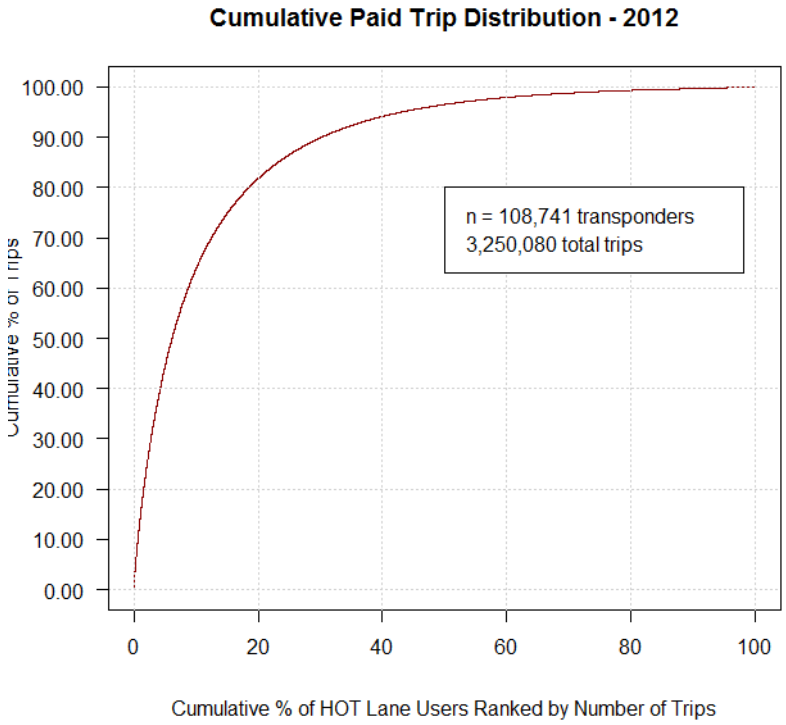


Figure 22: Paid Trip Distribution for 2012

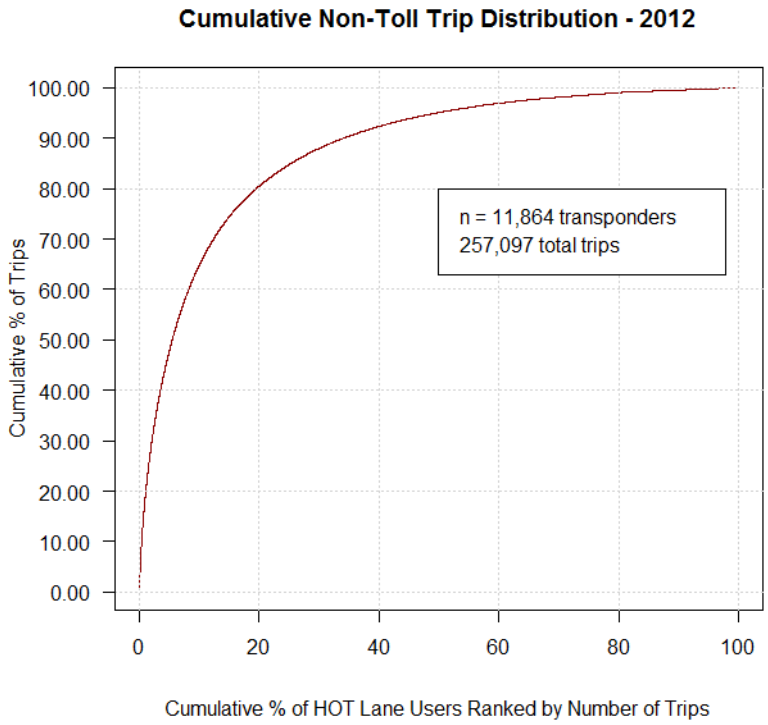


Figure 23: Non-Toll Trip distribution for 2012

Account Data: This stream describes the attributes of Peach Pass accounts, including all of the transponders and vehicles that are associated with those accounts. The three tables are updated daily to provide data concerning new accounts and changes to existing accounts. Those changes may include a vehicle's switch from toll mode to carpool mode or vice versa, a change in transponder status, or a change in account status. One important feature of the Account data stream is the lack of a joining element between the transponder data and the vehicle data. While both are associated with an account, there is no direct link between a vehicle and a transponder. As such, accounts that have multiple vehicles and multiple transponders do not indicate which vehicle is associated with which transponder. This many-to-many relationship complicates the process of linking transponders to vehicles and ultimately to the marketing demographic data. This issue is

explored further in other chapters, namely the Data Quality and Treatment chapter and the Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter.

Frequency: Delivered daily.

Data Fields:

Table 9: Base Account Data

Name	Description
AccountID	Unique Account Identifier
AccountType	Includes Personal, Corporate, Toll Exempt, Register by Plate, Non-Revenue, Emergency Non-Revenue, and Regular Post Paid
AccountStatus	Includes Active, Proposed, Pending to Close, Suspended, Closed, and Cancelled

Table 10: Account Transponder Data

Name	Description
AccountID	Unique Account Identifier
TransponderID	Unique Transponder Identifier. The table includes rows for each transponder associated with an account.
TransponderAgencyCode	Indicates whether the transponder was originally used for Georgia SR-400 (GA400) or for the I-85 Express Lanes (GSRTA)
TransponderStatus	Includes Active, Lost, Stolen, Inactive, No Balance, and Low Balance

Table 11: Account Vehicle Data

Name	Description
AccountID	Unique Account Identifier
PlateNumber	License plate of the vehicle. The table includes rows for each vehicle associated with an account.
PlateState	Registration state of the vehicle.
TollMode	Toll or Non-Toll
TollModeTimestamp	Timestamp of the switch to or from toll mode

Table 12 provides an overview of the Account data by type and status. These data are useful in separating out Peach Pass transponders that have been registered to corporate or toll-exempt accounts. As the behavior of these users is likely very different from that of other users, it may prove beneficial to model them separately. In the table below, account status type A refers to 'Active,' while I and P are 'Pending to Close' and 'Proposed,' respectively. Account types S, CC, and C indicate Suspended, Cancelled, and Closed accounts. An examination of the HOT lane trips from 2012 indicates that 1.2% of the trips were taken by Suspended, Cancelled, and Closed accounts. However, that figure is based on the population of trips for which a join to the Account data could be made. As mentioned above, the many-to-many relationship between transponders and vehicles narrows the scope of accounts for which this join is possible. As a result, the 1.2% figure reflects only those accounts that have a single vehicle and a single transponder. In addition, the table is based on account status results from August 2013; that may not have been the status of the account at the time the trip was made. The table shows that there are just over 270,000 accounts that are active or will be soon, and almost

100,000 accounts in the data set that are no longer active. Of the active accounts, most of them are personal. A non-trivial number, over 14,000, are corporate or toll-exempt. As mentioned above, these users may behave differently as the users themselves are not paying a toll.

Table 12: Accounts by Type and Status as of 8/21/2013

Account Type	All Accounts	Status = (A,I,P)	Status = (S, CC, C)
P (Personal)	347076	257480	89596
C (Corporate)	10197	8379	1818
TE (Toll Exempt)	5942	5727	215
RBP (Register By Plate)	4	2	2
NR (Non-Revenue)	13	3	10
ENR (Emergency Non-Revenue)	129	118	11
B (Regular Post-Paid)	528	333	195
Total	363889	272042	91847

The structure of the Account data provided by SRTA creates number of issues. Accounts that have multiple vehicles and multiple transponders cannot have those transponders connected to individual license plates. The frequency of this issue is further explored in other chapters: Data Quality and Treatment, and Potential Sample Bias in Paired vehicle Activity and Marketing Data. In addition, the "Toll Exempt" category within the account types is meant to include vehicles that will always have carpools, alternative fuel vehicles, and motorcycles. The Account data do not make any distinction between these three categories, however; they are all grouped together under the "Toll Exempt" umbrella. In developing future database structures for toll implementation, this dissertation will recommend that these and other data issues be avoided from the outset via improved database design.

Figure 24 combines data from the Trip stream and the Account stream to identify trips taken by transponders and accounts that both have 'active' status as of May, 2014.

The resulting cumulative trip distributions are plotted for 2012, 2013, and January through May of 2014. In each year, the top 10% of active transponders (by number of trips) take over 80% of the Express Lane trips. For both 2012 and 2014, those users take over 90% of the Express Lane trips.

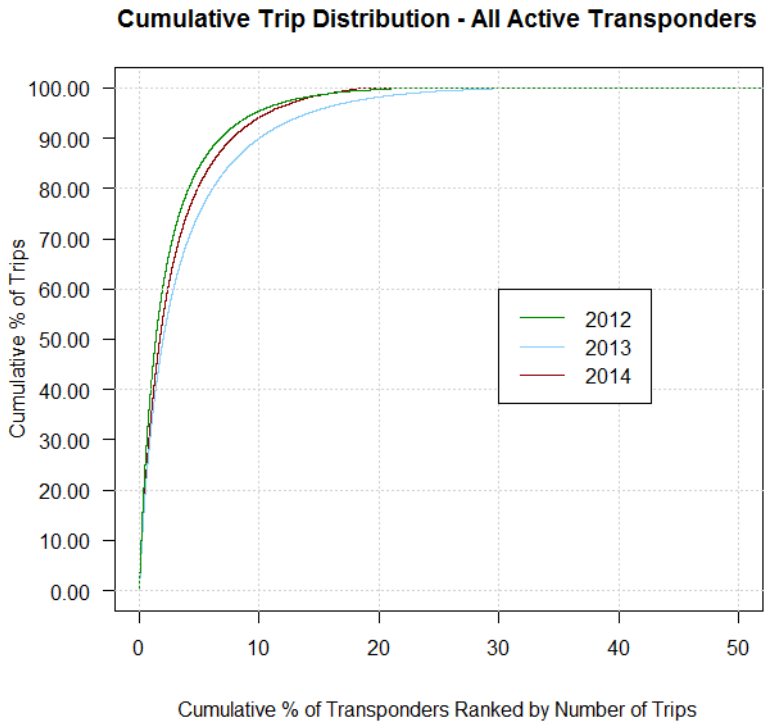


Figure 24: Cumulative Trip Distribution for all Active Transponders

Figure 25 shows the cumulative toll amount distribution for individual transponders ranked by the total amount of toll paid. The chart illustrates cumulative lines for calendar years 2012 and 2013. In each case, the top 10% of toll paying transponders paid over 75% of the total annual toll amount.

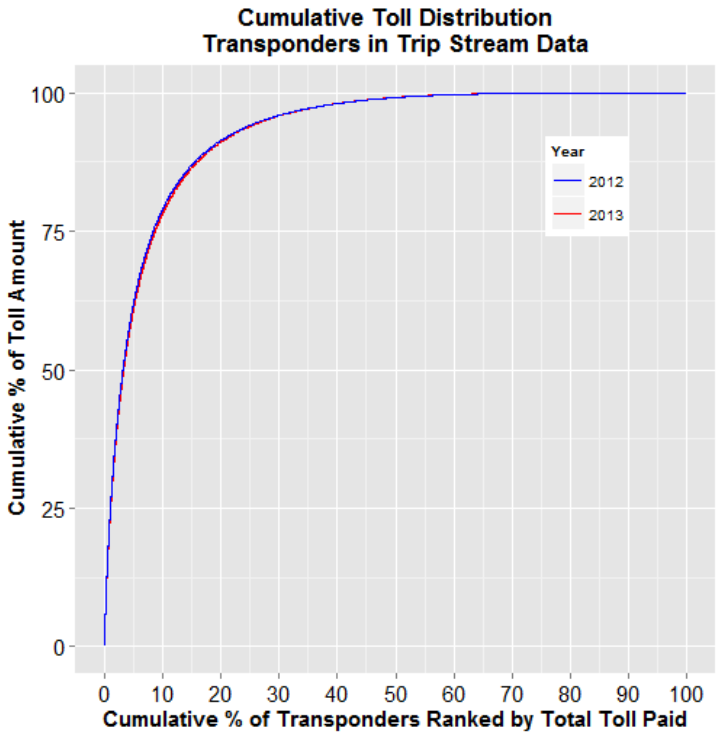


Figure 25: Cumulative Toll Distributions

Epsilon Marketing Data

The socio-economic data used in this dissertation come from credit reports and other data that are sanitized, processed, and packaged for sale by the marketing data firm Epsilon. These data include a host of variables at the household, neighborhood, and individual levels. This study will make use of the household and neighborhood level variables, as it is not possible to identify individual drivers in multi-person households with the available data. The full marketing data set originated from a list of license plates collected by researchers for the I-85 HOV-to-HOT conversion analysis project at Georgia Tech. This project involved the collection of license plates of users of the I-85 corridor. Researchers collected license plate data in the morning and afternoon peak periods at five sites along the corridor; this occurred four times a year, once each season, for two years. The first year of data collection occurred before the opening of the facility, while the second occurred immediately after operations began. Researchers identified frequent users of the corridor and targeted those license plates for the demographic data purchase (Khoeini, 2014). The complete set of data includes 349,134 records. The dataset was purchased and delivered to Georgia Tech on March 6, 2013. This section will provide an overview of the demographic variables in the marketing data set, and then examine potential correlation within the household data.

The full marketing demographic data set contains 130 data elements for each record. These elements include a combination of identifying variables (unique Epsilon identification number, address data, household name, variables that indicate what type of match was made and other elements that are related to Epsilon's internal processes), household and head of household demographic variables, and neighborhood demographic

variables. This section lists the full set of household and neighborhood variables; the individual variables and identifying data are not used in this dissertation's analysis. A complete table of all of the marketing variables is available in the Appendix.

Epsilon Data Coverage

A recent working paper by Khoeini (2013) compared the marketing data used in this dissertation with aggregate and disaggregate census data. The paper discussed the relative benefits of this marketing data: the price per household is significantly lower than that of survey data, and the data are updated more frequently (typically every three months). Unlike the extensive Atlanta Household Travel Survey from 2011 to which these data were compared, the marketing data do not provide travel or vehicle ownership information. While additional trip-related data would be useful for this dissertation, the SRTA lane use data provide a substantial amount already. One of the limitations of the marketing data identified by Khoeini is the issue of coverage, or the varying degrees of completeness for each observation. Many of the socioeconomic variables are not present for each household; for example, only 36% of the households have associated income data. Epsilon provides imputed data to fill in these gaps. While the imputation methods remain confidential, Khoeini reported that the inferred data closely matched the household travel survey data. Khoeini concluded that "the accuracy and coverage of marketing data are not as [good] as survey data," but that a "large enough sample of marketing data could potentially cancel out the errors across the user groups."

Table 13 provides an overview of the household variables in the Epsilon marketing dataset, along with the percentage of usable, non-blank records in each category. While many of the variables are self-explanatory, certain variables require

further description. The income variables, such as Household Income and Narrow Band Income, are ordinal variables for which each value is a range of household incomes. For this dissertation, the author used the midpoint of those income ranges. The table illustrates the varying rates of coverage among the household demographic variables; those relating to the physical houses themselves have the lowest rates of available data (property lot size, living area size, year of home construction, and home market value variables). In addition, the 'Home Market Value' variable is missing the required description of the potential values in the data dictionary. As currently presented, Home Market Value cannot be deciphered and used in analysis. The Household Age variable refers to the age of the head of the household, while the Occupation variable is described as the 'most prominent known profession of everybody in the household' (Epsilon Targeting, 2013).

Table 13: Epsilon Household Variables

Variable Name	Non-Blank Records	Blank Records	Coverage Amount
Living Area Square Feet	146,106	203,028	41.85%
Property Lot Size in Acres	117,367	231,767	33.62%
Year Home Built	144,803	204,331	41.48%
Home Market Value	257,259	91,875	73.68%
Household Income	348,435	699	99.80%
Dwelling Type	346,374	2,760	99.21%
Home Valuation Model	301,349	47,785	86.31%
Home Owner	344,503	4,631	98.67%
Household Education	348,251	883	99.75%
Household Marital Status	348,435	699	99.80%
Number of Adults	348,435	699	99.80%
Length of Residence	348,435	699	99.80%
Narrow Band Income	348,435	699	99.80%
Target Income	348,435	699	99.80%
Household Age	348,435	699	99.80%
Presence of Children	348,435	699	99.80%
Household Size	348,435	699	99.80%
Occupation	306,766	42,368	87.86%

Table 14 below presents an overview of the neighborhood demographic variables and the number of non-blank records in each category. Here the coverage rates are equivalent across all variables. For the three variables in which it is used, 'Average CMV' refers to average commercial market value. These variables were initially ignored in the analytical process, however later examination indicated that they may contribute to choice modeling. This is discussed in further detail in Chapter 12, the Modeling Extensions chapter of this dissertation.

Table 14: Epsilon Neighborhood Variables

Variable Name	Non-Blank Records	Blank Records	Coverage Amount
Percent of Households Owning a Registered Passenger Car	348,830	304	99.91%
Percent of Households Owning a Registered New Passenger Car	348,830	304	99.91%
Percent of Households Owning a Registered Truck	348,830	304	99.91%
Percent of Households Owning a Registered New Truck	348,830	304	99.91%
Percent of Households Owning a Registered Motorcycle	348,830	304	99.91%
Average CMV in Thousands for all New and Used Registered Vehicles	348,830	304	99.91%
Average CMV in Thousands for all New and Used Registered Cars	348,830	304	99.91%
Average CMV in Thousands for all New and Used Registered Trucks	348,830	304	99.91%
Percent of Households Owning a Registered Motor Home	348,830	304	99.91%

Selected Variable Distributions

Figure 26 through Figure 31 illustrate distributions of a small subset of the demographic variables in the full Epsilon dataset. Figure 26 presents the household income distribution with varying column widths to represent the differences in categorical ranges. The maximum income cutoff is set at $300,000 for the purposes of visual representation; that category actually includes all household incomes over $250,000. A plurality of households, nearly 25%, fall into the $50,000-$74,999 range. The two lowest income categories, those from $0-14,999 and $15,000-19,999, make up 8.54% of the data set.

Very few households, less than 1% of the total, have annual incomes exceeding $200,000. The median income in this sample, $62,500, exceeds the Census Bureau American Community Survey five-year median estimates for the City of Atlanta ($46,631) and the state of Georgia ($49,179) (U.S. Census Bureau, 2013). Further discussion of the Census data can be found in the Connecting SRTA Data to Epsilon Data and Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapters of this dissertation.

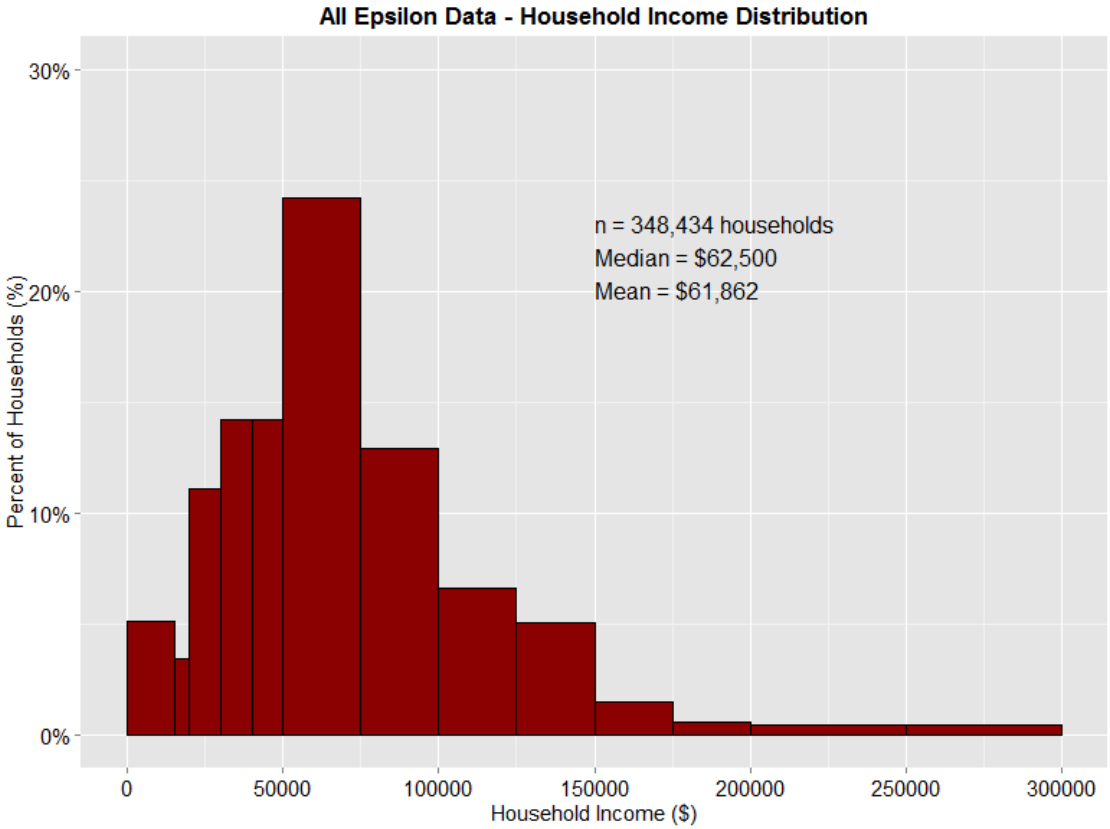


Figure 26: Household Income Distribution in All Epsilon Data

Figure 27 presents the distribution of household education levels in the full Epsilon dataset. The majority of households, nearly 70%, have completed some or all of an undergraduate degree. Over 27% of the households completed only a high school

education. Only 2.33% of corridor users in the data set did not finish high school. However, graduate level education is even more rare in this sample (0.92%).

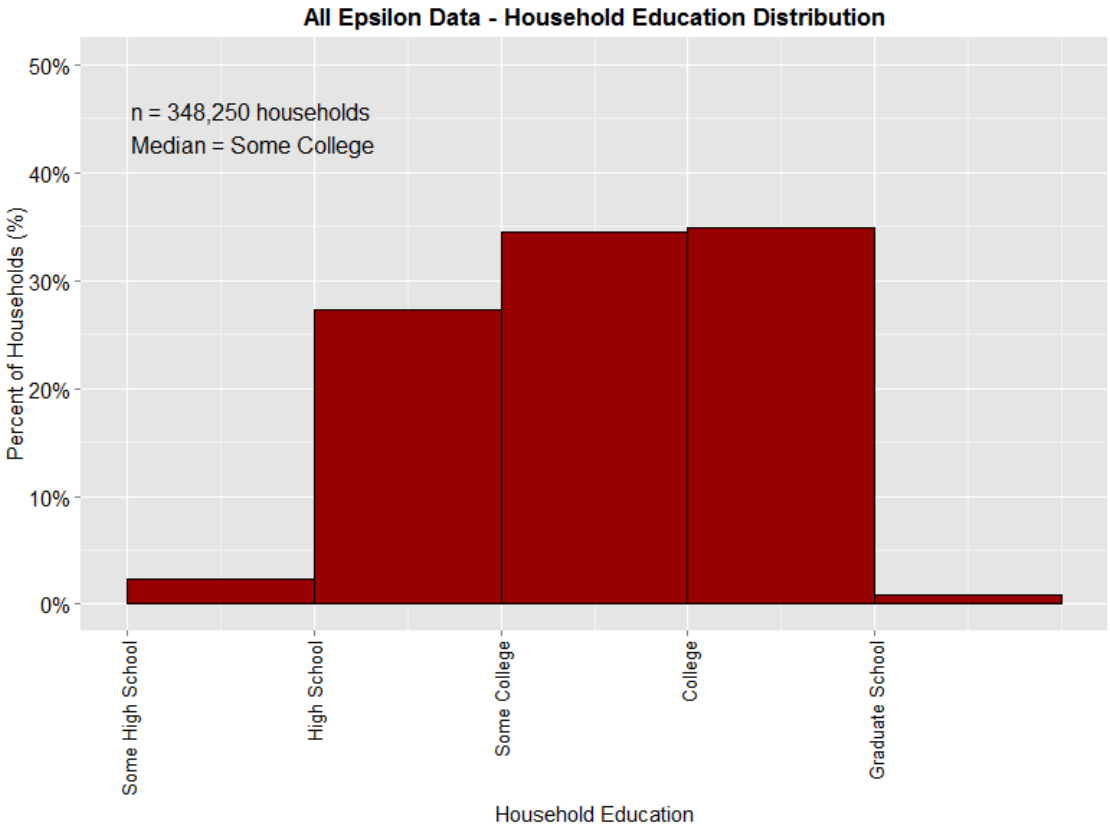


Figure 27: Household Education Distribution in All Epsilon Data

Figure 28 shows the distribution of the heads of household ages. Here nearly 35% of households have a household head in the 35-44 year old age range. Households with head of households under 25 years old make up less than 1% of the total sample, while those over 75 years old comprise 3.14% of the total. The second largest category is the 45-54 years old range; this segment makes up nearly 24% of the households.

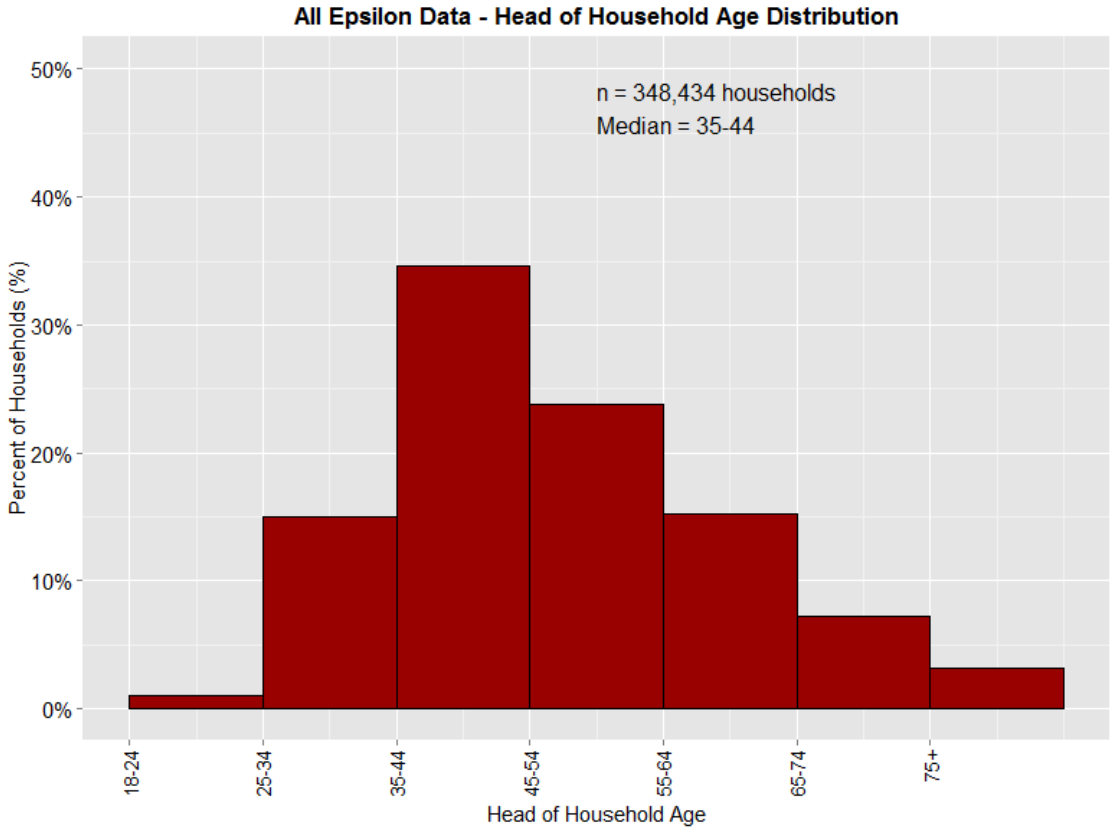


Figure 28: Head of Household Age Distribution in All Epsilon Data

Figure 29 shows household sizes in the Epsilon data. Nearly 37% of all households consist of a single individual; two-person households make up 25.3% of the total. The final category includes households with nine or more persons; these make up 0.69% of the full Epsilon dataset.

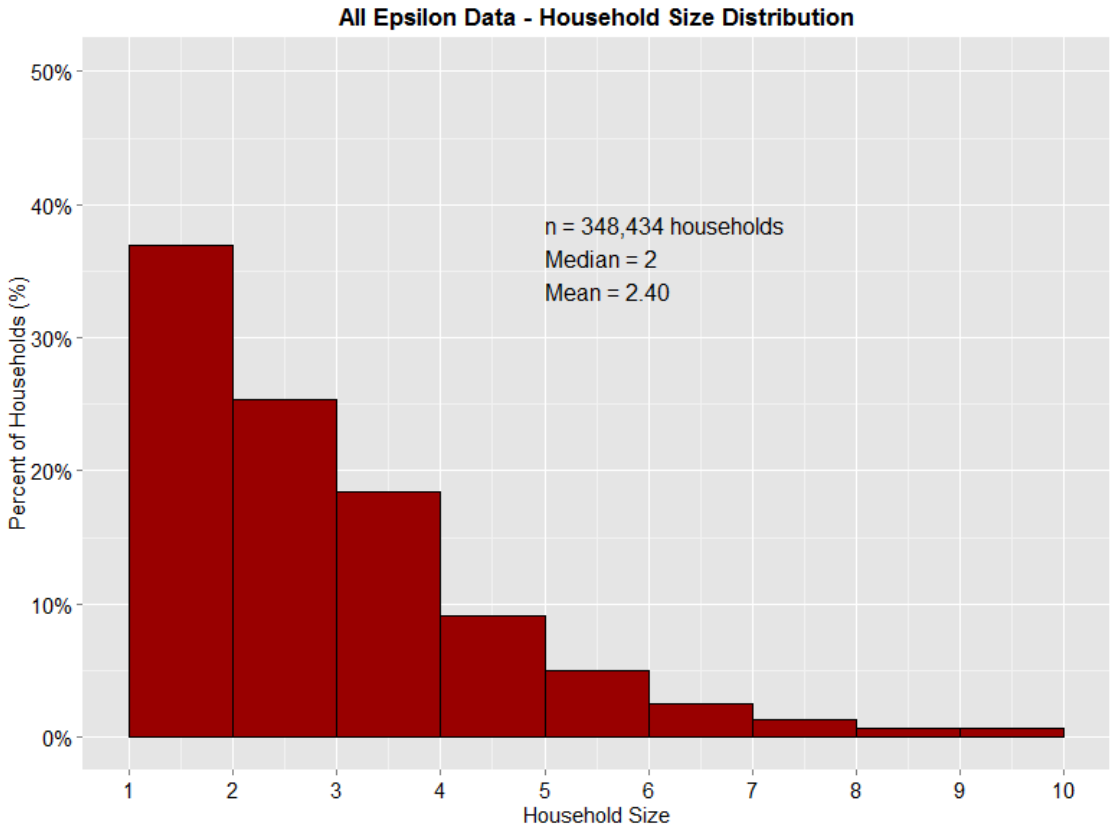


Figure 29: Household Size Distribution in All Epsilon Data

The Epsilon data also categorize households as renters or owners of their homes, and assign a 'definite' or 'probable' rating to the results. Figure 30 presents the distribution of these owner or renter assignments. Over 50% of the sample consists of what Epsilon deems to be 'definite owners,' while 'definite renters' make up only 0.55% of the total. 'Probable owners' dwarf the share of 'probable renters' too; the former category has 30.4% of the households, while the latter has approximately 15%.

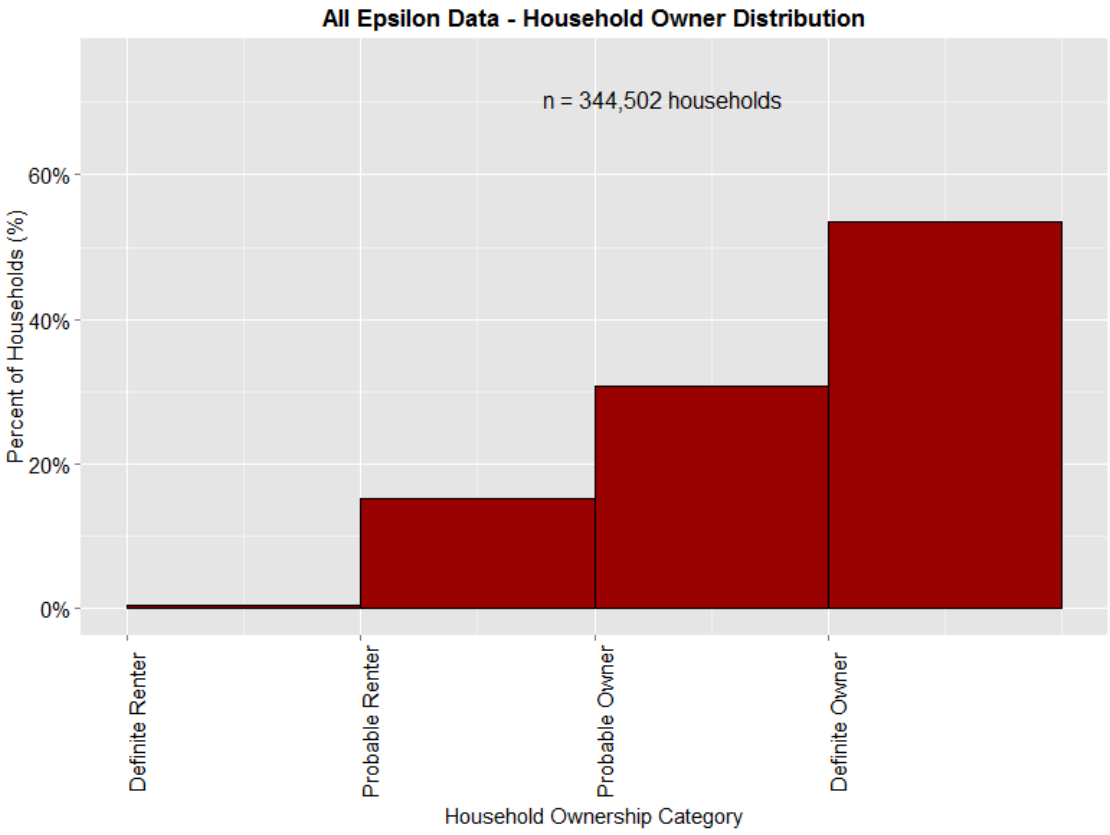


Figure 30: Household Ownership Distribution in All Epsilon Data

The final demographic variable presented here is the dwelling type of the households. Epsilon categorizes household dwelling types as SFDU (single family dwelling unit), MFDU (multi-family dwelling unit), business, CMRA (commercial mail receiving agency), condo, and mobile home. The vast majority of households, nearly 87%, live in single family dwelling units. At 8.86% of observations, multi-family dwelling unit households make up approximately one tenth of the single-family unit count. A trivial amount of records fall into the business or mobile home categories; each contains less than 1% of the total households. The marketing firm does not explain why condominiums are listed separately from multi-family dwelling units, but they comprise less than 3% of all households.

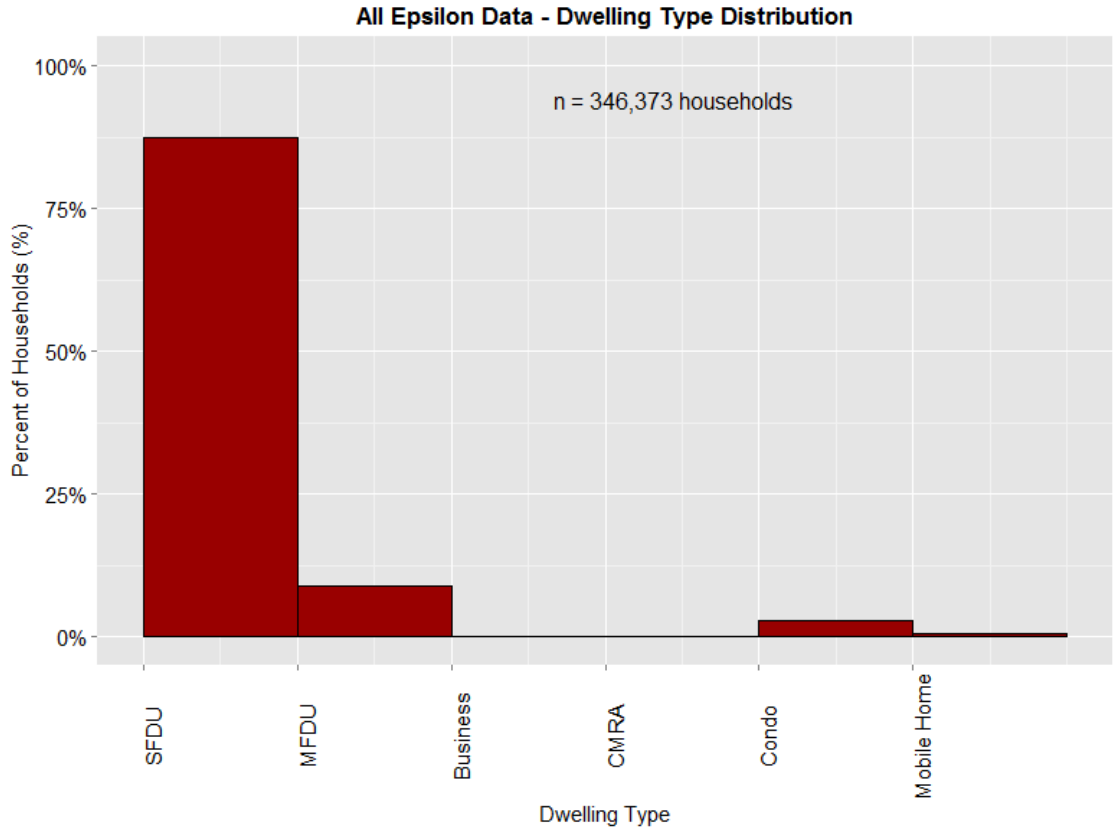


Figure 31: Dwelling Type Distribution in All Epsilon Data

Correlation within Demographic Data

The correlation matrix shown in Figure 32 was computed using Pearson's method with pairwise complete observations. The correlation coefficient values presented in the matrix are also color-coded along a gradient to represent the level of correlation. The dark green shown in the diagonal 1-values represents the highest level of positive correlation; high levels of negative correlation are presented in dark red. Yellow values represent positive correlation coefficients that are low in magnitude. Similarly, orange values highlight negative correlation values with small magnitudes. The intersections between the four highlighted variables are also surrounded in thick black borders for easier identification. Four of the variables have been assigned colors within the left-most name column to better identify their positions in the top row. These four variables have also been used in the preliminary analyses that are presented later in the dissertation: household income, household size, head of household age, and household education level.

The results of the correlation matrix include both expected and unexpected values. The household income variable, represented in blue, is positively correlated with most of the remaining variables. The largest positive coefficients can be seen with household education, living area size, and home ownership status. The largest correlation coefficient is found between the two different income variables, as may be expected. The coefficient estimated for household income and property lot size is negative, but the magnitude is very small. The other negative coefficients, between income and dwelling type and income and marital status, are an artifact of the manner in which the dwelling type and marital status variables are coded. A dwelling type of value 1 is a single-family

dwelling unit, which represents the vast majority of households in the Epsilon data. Similarly, a marital status value of 1 represents marriage, while 2 is single. In terms of the other highlighted variables, income is more strongly correlated with household size than with age.

Household education is, again, positively correlated with income, with a correlation coefficient of 0.46. The correlation values with head of household age and size are also positive, but much smaller in magnitude. Education is positively correlated with living area size with a coefficient of 0.30, but negatively correlated with property lot size with a coefficient of -0.30. This represents the largest coefficient value, either positive or negative, among the property lot size values.

The head of household age variable is most highly correlated with just two variables: home ownership and length of residence. The coefficient value between household age and income is weakly positive, perhaps due to seniors whose earning power has decreased. Note that income data do not speak to wealth and availability of money, and more detail may be found by examining income and employment status in conjunction. A similar effect may explain the weak correlations between household age and household size, as well as household age and number of children: older households see their children depart and form new households.

The final highlighted variable, household size, is not surprisingly strongly positively correlated with number of children, number of adults, and presence of children. Other positive but weaker correlation coefficients appear with household income, home ownership, and length of residence. The strongly negative correlation between household

size and marital status is again explained by the coding of the marital status variable, in which a value of 1 indicates marriage.

The variable correlations that are illuminated by this correlation matrix are largely expected. They include strongly positive relationships between income and education, income and household size, and marital status and household size. Perhaps the most surprising results are those related to property lot size, in particular its strongly negative relationship with household education. These results will inform future analyses that involve demographic data, particularly those in which multiple demographic elements are included.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Ethnic Group	1	0.051275	0.038503	0.150648	0.190378	-0.03249	0.106702	0.051519	-0.11843	0.076881	0.066092	0.123496	0.04326	0.071629	0.086273
2. Number of Children	0.051275	1	0.076066	0.008233	0.121037	-0.05592	0.118347	0.007527	-0.16528	0.06385	0.061416	0.075328	-0.1021	NA	0.837566
3. Living Area Square Feet	0.038503	0.076066	1	0.10843	0.477166	-0.20886	-0.00357	0.301528	-0.19133	0.145882	-0.04568	0.580409	0.025834	0.153089	0.188639
4. Property Lot Size in Acres	0.150648	0.008233	0.10843	1	-0.00091	-0.08828	-0.06949	-0.30421	-0.13691	0.16307	0.176604	-0.08046	0.169458	0.02174	0.104358
5. Household Income	0.190378	0.121037	0.477166	-0.00091	1	-0.11154	0.396316	**0.459989**	-0.29063	0.276934	0.308781	0.665654	**0.189761**	0.215315	**0.304519**
6. Dwelling Type	-0.03249	-0.05592	-0.20886	-0.08828	-0.11154	1	-0.28088	0.033406	0.133021	-0.10039	-0.10568	-0.17944	-0.08312	-0.10232	-0.11743
7. Home Owner	0.106702	0.118347	-0.00357	-0.06949	0.396316	-0.28088	1	0.083576	-0.2772	0.389081	0.454147	0.399027	0.355374	0.227611	0.359763
8. Household Education	0.051519	0.007527	0.301528	-0.30421	**0.459989**	0.033406	0.083576	1	-0.00643	0.026079	0.029669	0.479948	**0.026008**	-0.00547	**0.011665**
9. Marital Status	-0.11843	-0.16528	-0.19133	-0.13691	-0.29063	0.133021	-0.2772	-0.00643	1	-0.50161	-0.30464	-0.27224	-0.21264	-0.36515	-0.46112
10. Number of Adults	0.076881	0.06385	0.145882	0.16307	0.276934	-0.10039	0.389081	0.026079	-0.50161	1	0.428299	0.264624	0.3245	0.31085	0.730325
11. Length of Residence	0.066092	0.061416	-0.04568	0.176604	0.308781	-0.10568	0.454147	0.029669	-0.30464	0.428299	1	0.197036	0.450807	0.231095	0.374375
12. Household Income (narrow band)	0.123496	0.075328	0.580409	-0.08046	0.665654	-0.17944	0.399027	0.479948	-0.27224	0.264624	0.197036	1	0.024012	0.23058	0.287165
13. Head of Household Age	0.04326	-0.1021	0.025834	0.169458	**0.189761**	-0.08312	0.355374	**0.026008**	-0.21264	0.3245	0.450807	0.024012	1	-0.09283	**0.107147**
14. Presence of Children	0.071629	NA	0.153089	0.02174	0.215315	-0.10232	0.227611	-0.00547	-0.36515	0.31085	0.231095	0.23058	-0.09283	1	0.656047
15. Household Size	0.086273	0.837566	0.188639	0.104358	**0.304519**	-0.11743	0.359763	**0.011665**	-0.46112	0.730325	0.374375	0.287165	**0.107147**	0.656047	1

Figure 32: Epsilon Demographic Data Correlation Matrix

CHAPTER 4

DATA PROCESSING

The large amount of raw data delivered by SRTA and available in the Epsilon marketing dataset required a great deal of processing to be put in a usable form. This chapter will describe the various stages of the process that converts the abundant raw data into processed data usable for analysis. The first section describes the method of aggregating individual vehicle detections into vehicle trips. The second section provides an overview of the set of constructed trips. The third discusses the average travel time calculations and presents a selection of the results. The method of counting detected transponders along the corridor is presented in section four. The final section provides an overview of the process that joins constructed trips to the other elements of the SRTA data stream.

Building Trips from Disaggregated Detections

The Vehicle detection stream provided by SRTA and described in the Data Sources chapter delivers disaggregated data from each of the RFID tag readers along the corridor. For the purposes of this research, disaggregate tag read data need to be combined into vehicle trips. This section describes how the individual detections were aggregated into trips and provides an overview of the trips in the resulting dataset.

The algorithms that combine individual detections into vehicle trips begin by ordering chronologically all of the detections for a given Peach Pass transponder for a given day. The first detection in the resulting ordered list is identified as the start of the first trip. The script then loops through the remaining detections and either adds them to

the existing trip or creates a new trip. A detection is added to an existing trip if it meets three criteria: 1) the detection occurs within a specified time interval of the previous detection, 2) the detection is in the same direction as the previous detection (northbound or southbound), and 3) the detection occurs downstream of the previous detection. For this work, researchers used a time threshold of fifteen minutes between detections. If two detections occurred more than fifteen minutes apart, the second detection triggered the start of a new trip. The purpose of this is to break apart trip chains which exit the corridor and then re-enter soon after. A new trip is also triggered when the other two criteria are not met. The script continues to cycle through all of the day's detections for the given transponder in this manner.

Trips with speeds greater than 100 miles per hour were excluded from the data. This filter was implemented due to detections that were perceived to be mistimed or misreported, resulting in unreasonable trip speeds. Very few of the generated trips met this criterion; less than 0.1% of trips on any given day. Additionally, trips that had speeds of 0 mph (due to two detections being reported at the same time) were also removed from the data set. This screening step also eliminated less than 0.1% of the trips on any given day. Trips that started or ended on SR-316 were constructed, but there are no General Purpose tag readers on that branch of the Express Lanes. This made it impossible to compare conditions for the two lane types on SR-316 with the given data, and so these trips will not be included in the analyses. For each built trip, the elements that are collected and stored are listed and described below in Table 15.

Table 15: Fields in Constructed Trips

Field Name	Description
transponderId	Unique Peach Pass identifier
Year	Year of trip
Month	Month of trip
Date	Date of trip
hour	Hour of trip
fiveMin	Five minute interval of trip
fifteenMin	Fifteen minute interval of trip
startTime	Trip start time
endTime	Trip end time
travelTime	Duration of trip (seconds)
direction	Northbound or southbound
startLane	LaneID of first detection
endLane	LaneID of last detection
startGantry	Gantry name at first detection
endGantry	Gantry name at last detection
startSegment	Corridor segment at first detection
endSegment	Corridor segment at last detection
gpStartTime	Start time of GP portion of trip
gpEndTime	End time of GP portion of trip
gpStartLane	LaneID of first detection in GP portion of trip
gpEndLane	LaneID of last detection in GP portion of trip
gpStartGantry	Gantry name at first detection in GP portion of trip
gpEndGantry	Gantry name at last detection in GP portion of trip
gpEquivalentSection	Section number of similar trip in HOT lane
gpTravelTime	Duration of GP portion of trip (seconds)
gpDistanceft	Distance of GP portion of trip (feet)
gpSpeed	Speed of GP portion of trip (mph)
htStartTime	Start time of HOT portion of trip
htEndTime	End time of HOT portion of trip
htStartLane	LaneID of first detection in HOT portion of trip
htEndLane	LaneID of last detection in HOT portion of trip
htStartGantry	Gantry name at first detection in HOT portion of trip
htEndGantry	Gantry name at last detection in HOT portion of trip
htStartSegment	Corridor segment at first HOT detection
htEndSegment	Corridor segment at last HOT detection
htSection	SRTA Section number of trip
htTravelTime	Duration of HOT portion of trip (seconds)
htDistanceft	Distance of HOT portion of trip (feet)
htSpeed	Speed of HOT portion of trip (mph)
numberOfDetections	Total number of detections in the trip
misdetections	Number of detections with misreported times or locations
distanceft	Distance between the gantries that reported the first and last detections (feet)
distancemi	Distance between the gantries that reported the first and last detections (miles)
speed	Speed of entire trip (mph)
hotUse	Flag indicating HOT detections occurred
mixedTrip	Flag indicating HOT and GP detections occurred
segmentOP	Flag indicating vehicle was detected in the Old Peachtree segment
segmentPH	Flag indicating vehicle was detected in the Pleasant Hill segment
segmentIT	Flag indicating vehicle was detected in the Indian Trail segment
segmentJC	Flag indicating vehicle was detected in the Jimmy Carter Boulevard segment
segment285	Flag indicating vehicle was detected in the I-285 segment

In this list, the “gp start lane/gantry” and “gp end lane/gantry” entries refer to the first GP segment of the trip. That is, if the vehicle is detected in the GP lane, and then switches to the HOT lane, and then back to the GP lane, only the first GP lane detections will appear in the output. Additionally, the script only outputs trips that consist of more than one detection. A sample of the constructed trip output with one trip from January 1, 2014 is shown here in Figure 33.

```
1  transponderId,year,month,date,hour,fiveMin,fifteenMin,startTime,endTime,travelTime,direction,startLane,endLane,
2  startGantry,endGantry,startSegment,endSegment,gpStartTime,gpEndTime,gpStartLane,gpEndLane,gpStartGantry,gpEndGantry,
3  gpEquivalentSection,gpTravelTime,gpDistanceft,gpSpeed,htStartTime,htEndTime,htStartLane,htEndLane,htStartGantry,
4  htEndGantry,htStartSegment,htEndSegment,htSection,htTravelTime,htDistanceft,htSpeed,numberOfDetections,misdetections,
5  distanceft,distancemi,speed,hotUse,mixedTrip,segmentOP,segmentPH,segmentIT,segmentJC,segment285
6  00342838,2014,1,1,0,55,45,2014-01-01 00:57:50,2014-01-01 01:02:44,294,NB,170506,170520,GPN2,GPN4,GPN,GPN,
7  2014-01-01 00:57:50,2014-01-01 01:02:44,170506,170520,GPN2,GPN4,8,294,28195,
8  65.3872912801,None,None,None,None,None,None,,,,0,0,0,3,0,28195.0,5.33996212121,65.3872912801,0,0,0,0,1,1,1
```

Figure 33: Sample Constructed Trip Output

Below is an example of a series of RFID detections for a specific transponder that includes both HOT and GP lane detections. The vehicle is initially detected in the HOT lane. In the middle of its HOT lane trip, it is detected by a GP lane detector. After that detection, the vehicle continues in the HOT lane for the remainder of its trip.

Table 16: Sample Trip with HOT and GP Detections

Transponder	Lane	Gantry
00040787	170000	285N1
Trip Date	170012	285N2
April 20, 2012	170026	285N3
	170036	285N4
	170048	285N5
	170061	285N6
	170086	JCN1
	170099	JCN2
	170112	JCN3
	170125	JCN4
	170522	GPN3
	170137	JCN5
	170150	ITN1
	170163	ITN2
	170175	ITN3
	170187	ITN4
	170200	ITN5
	170212	PHN1
	170224	PHN2
	170238	PHN3
	170250	PHN4
	170262	PHN5
	170274	PHN6
	170288	PHN7

The corresponding constructed trip output for these data is shown here in Figure 34:

```
transponderId,year,month,date,hour,fiveMin,fifteenMin,startTime,endTime,travelTime,direction,startLane,endLane,
startGantry,endGantry,startSegment,endSegment,gpStartTime,gpEndTime,gpStartLane,gpEndLane,gpStartGantry,gpEndGantry,
gpEquivalentSection,gpTravelTime,gpDistanceft,gpSpeed,htStartTime,htEndTime,htStartLane,htEndLane,htStartGantry,
htEndGantry,htStartSegment,htEndSegment,htSection,htTravelTime,htDistanceft,htSpeed,numberOfDetections,misdetections,
distanceft,distancemi,speed,hotUse,mixedTrip,segmentOP,segmentPH,segmentIT,segmentJC,segment285
00040787,2012,4,20,22,0,0,2012-04-20 22:02:43,2012-04-20 22:11:49,546,NB,170000,170288,28501,PH07,285N,PHN,2012-04-20
22:06:44,2012-04-20 22:06:44,170522,170522,GPN3,GPN3,NA,0,0,0,2012-04-20 22:02:43,2012-04-20
22:11:49,170000,170288,28501,PH07,285N,PHN,4,546,58510,73.0644355644,24,0,58510.0,11.0814393939,73.
0644355644,1,1,0,1,1,1,1
```

Figure 34: Sample Built Trip with Mixed Detections

This record indicates that the trip began at the first northbound HOT gantry in the I-285 segment and ended at SR-316E. Note that the GPN3 detection which occurred in the middle of the trip did not break up the HOT trip that was reported by the script. The trip-building algorithm allows for Express Lane trips to continue after a single General Purpose lane detection. This resembles the logic of SRTA's trip building, which also ignored that GP detection and reported the same start- and end-points of the HOT trip (section 4 starts at 285 northbound and ends at Pleasant Hill northbound). Figure 35 shows the corresponding trip record in the SRTA Trip summary stream, which also has the same start and end time, along with the same transponder identifier.

tripId	sectionId	tollAmount	tollMode	transponderId	transponderAgencyCode
1961697	4	0.12	TOLL	00040787	GSRTA

vehiclePlateNumber	vehiclePlateState	tripEntryTime	tripEntryExit
B[redacted]	GA	2012-04-20 22:02:43	2012-04-20 22:11:49

Figure 35: Corresponding Sample SRTA Trip

Characteristics of Constructed Trip Dataset

The full set of constructed trips was built from the individual RFID vehicle detections that were generated with the opening of the facility on October 1, 2011. Table 17 shows the number of total trips, Express Lane-only trips, GP lane-only trips, and mixed trips for 2012, 2013, and 2014. The total number of trips by Peach Pass equipped vehicles has been increasing by at least one million per year. The rates of toll lane trip-taking vary each year, but not in a consistent direction: while the proportion of general purpose-exclusive trips increased slightly from 2012 to 2013, it decreased again in 2014.

Table 17: Number of Constructed Trips by Year

	2012	2013	2014
Total trips by Peach Pass -equipped vehicles	11,188,848	13,903,170	15,250,085
HOT-Only trips	1,540,232 (13.8%)	1,683,636 (12.1%)	2,169,130 (14.2%)
GP-Only trips	7,059,956 (63.1%)	8,854,212 (63.7%)	9,480,632 (62.2%)
Mixed trips	2,588,660 (23.1%)	3,365,322 (24.2%)	3,600,323 (23.6%)

Figure 36 shows the distribution of RFID detections per trip for HOT-exclusive trips for January of 2013. A single month was selected as an example because of the large amount of data in each month's worth of constructed trips. The maximum number of detections possible is 35, as 35 RFID detectors span the length of the Express Lanes. A plurality of trips, nearly 25%, consist of 24 detections. These 24 detections make up 69% of the 35 detector total. Vehicles may trigger every RFID gantry they pass under during corridor trips; they may also miss a tag read if a scanner is out of commission. The trip-building scripts used in this dissertation account for such occurrences.

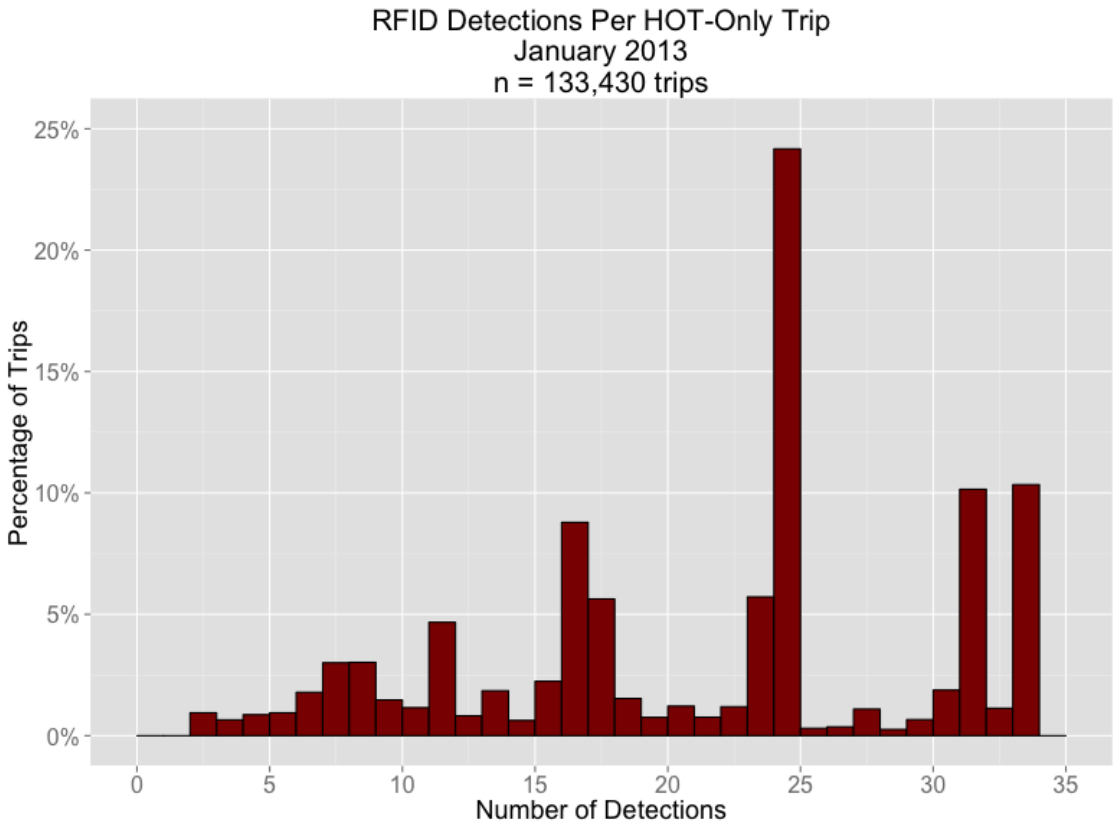


Figure 36: RFID Detections per HOT-Only Trip

Figure 37 illustrates the distribution of RFID detections per GP-only trip in January, 2013. As vehicles detected at just a single gantry are not included in the constructed trip set, the minimum number of detections in the constructed trip set is two. Roughly 25% of the trips occur across three general purpose lane vehicle detection gantries. Trips with seven detections only occur in the northbound direction, as there are only six general purpose vehicle detectors in the southbound direction. As is the case with the Express Lane detections, relatively few trips include a detection at each gantry along the corridor. It is also the case that GP-lane RFID scanners are adjacent to HOT-lane scanners. In addition to scanners missing detections, they may also double-count transponders by detecting them in both lane types.

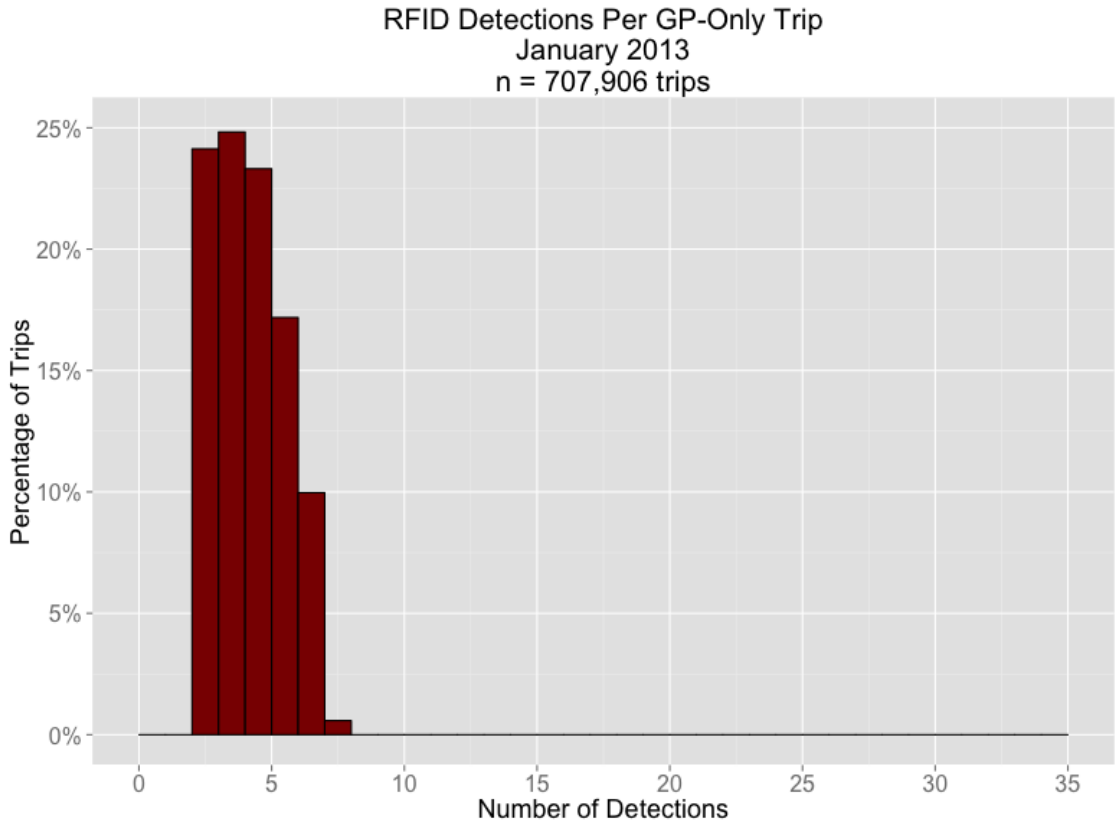


Figure 37: RFID Detections per GP-Only Trip

Figure 38 below illustrates the kernel density distribution of constructed trip speeds for all trips in March, 2012. The colored lines represent mixed trips (blue; those that use both the GP and HOT lanes), GP-only trips (red), and HOT-only trips (green). This figure includes all hours of the day and all days of the week. Note that the GP distribution is dominated by the effect of off-peak trips while the majority of HOT trips occur in the peak periods; Figure 43 presented later will illustrate this effect. All of the figures below were generated from the constructed trip data set with the detection interval set to 900 seconds between gantry detections.

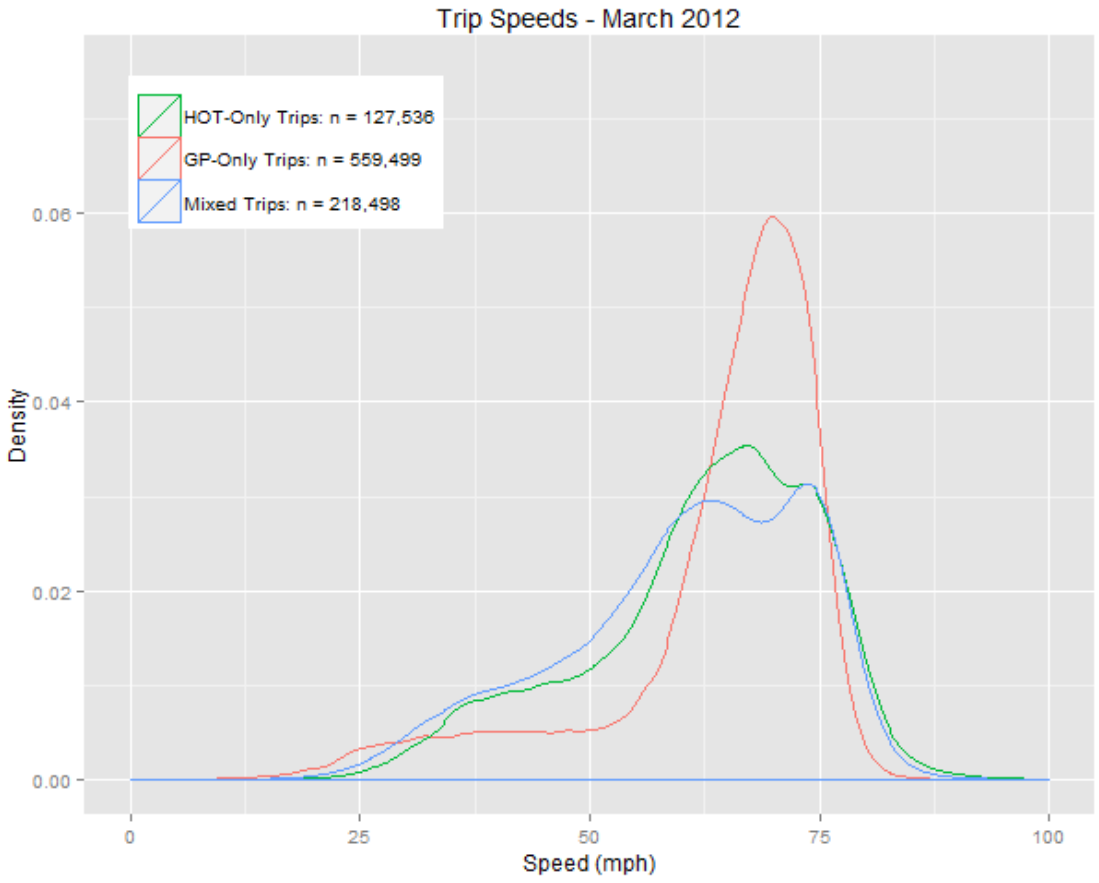


Figure 38: Speed Distribution for March 2012 Constructed Trips

Figure 39 and Figure 40 illustrate the kernel density distributions for peak-period trips in March, 2012. Unlike Figure 38, weekends were excluded from these distributions (along with off-peak hours). Here the benefits of the Express Lane are more pronounced: HOT speeds are more concentrated at the higher end of the speed distribution, and fewer Express Lane trips are observed around the 25-30mph range where a plurality of GP lane trips take place. This effect is even more pronounced in the northbound PM peak period trips shown in Figure 40. Here the higher-speed peak includes far more of the total trips, while almost no trips occur at the lower end of the distribution. In both cases, the 'mixed' trips, which traverse both lane types, behave more like toll lane-only trips than general purpose lane-only trips. Figure 41 and Figure 42 present these speed distributions as cumulative distribution functions for the southbound and northbound trips respectively.

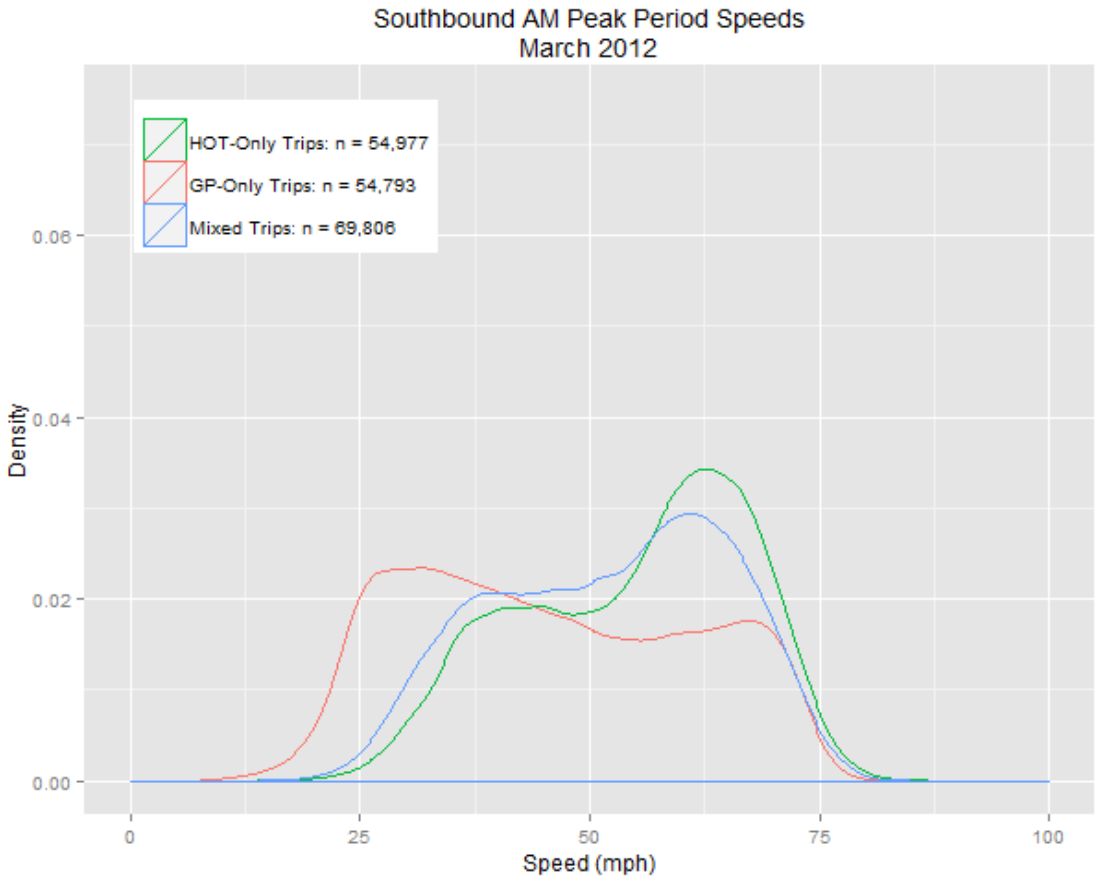


Figure 39: Speed Distribution for March 2012 SB AM Built Trips

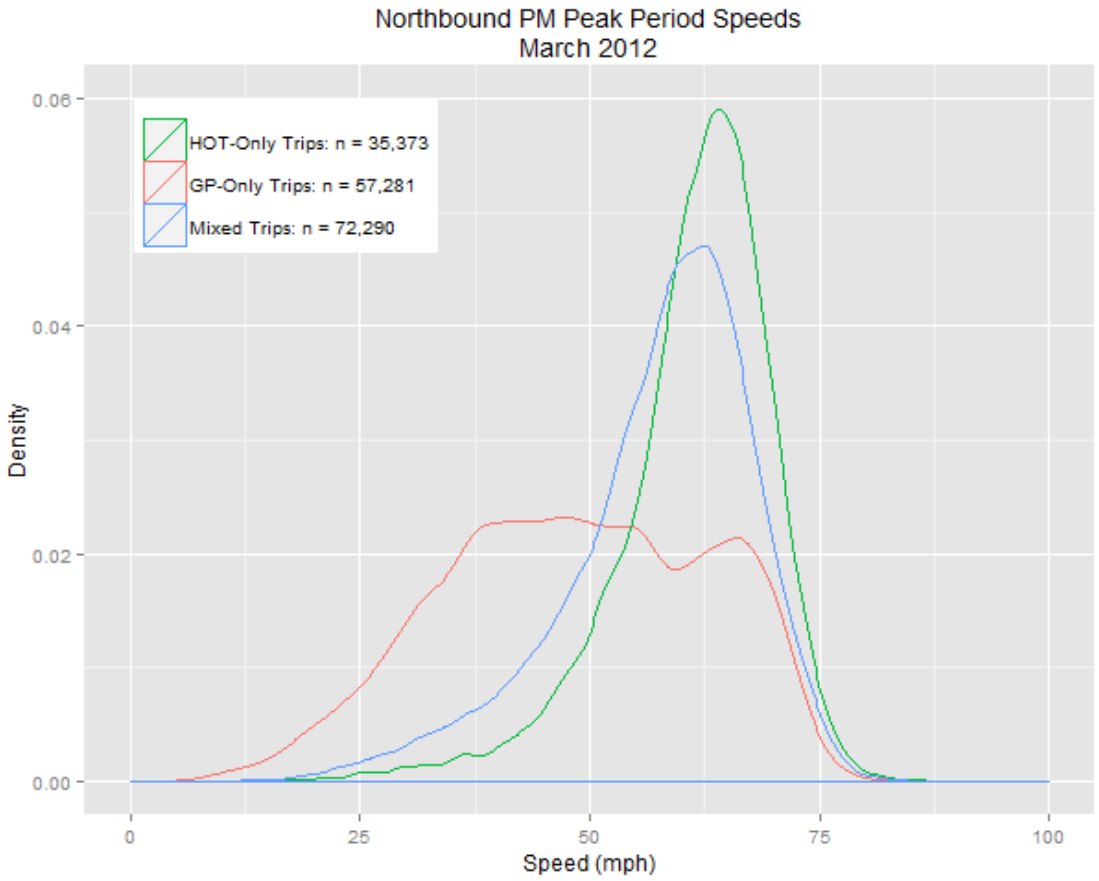


Figure 40: Speed Distribution for March 2012 NB PM Built Trips

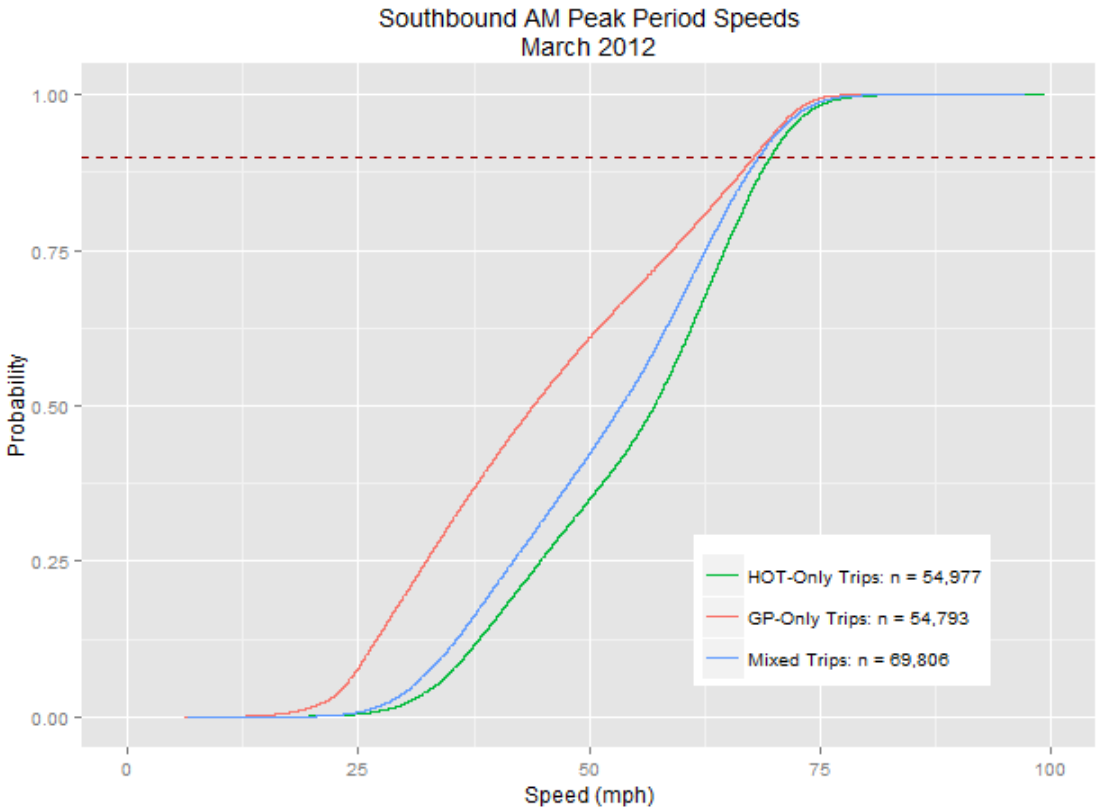


Figure 41: Speed CDF for March 2012 SB AM Built Trips

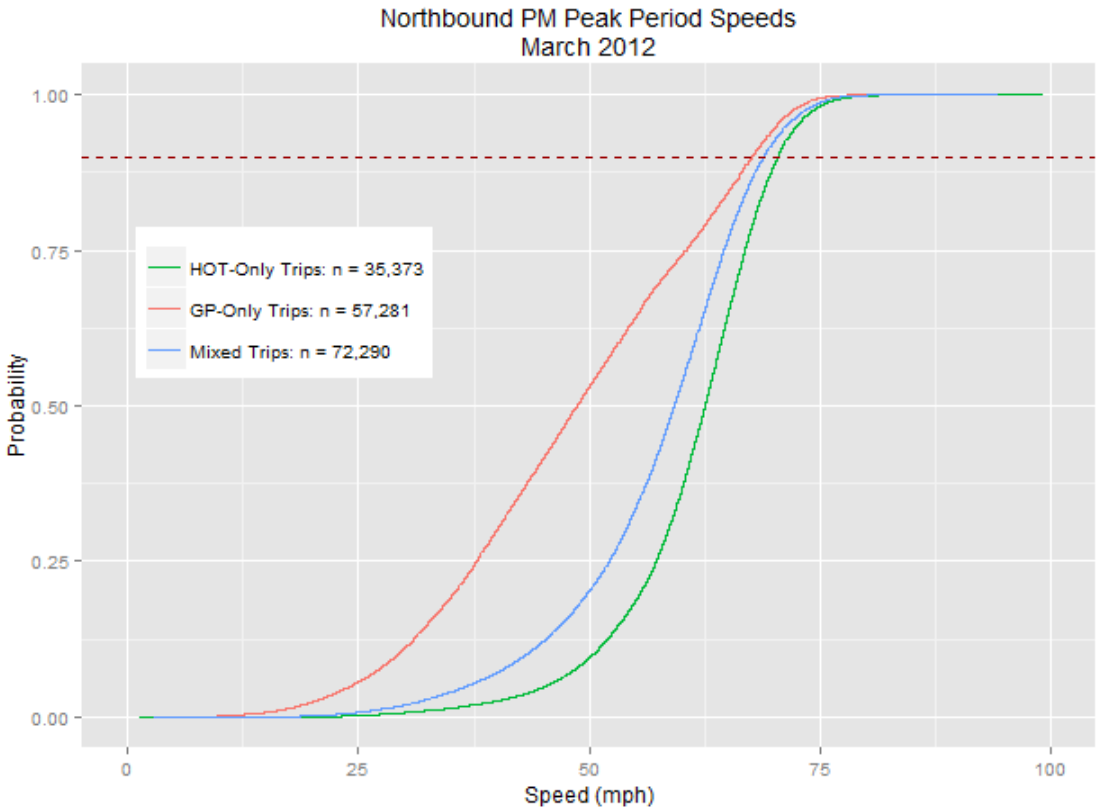


Figure 42: Speed CDF for March 2012 NB PM Built Trips

Figure 43 illustrates the kernel density distribution of off-peak trips for both lane types. Peak periods are defined as 6-10:00 AM in the southbound direction and 3:00-7:00 PM in the northbound direction. Here the distributional center of the GP-lane trips is lower than that of the toll lane trips, though in both cases the majority of trips are taken at high speeds. Note the discrepancy in trip counts between the HOT-only and GP-only trips: the count of off-peak GP trips is an order of magnitude higher than that of off-peak HOT trips. These trips likely dominate the GP-lane speed distribution in Figure 38, which shows the unpriced lanes carrying more high speed trips than the HOT lanes. This figure illustrates the benefits provided by the Express Lanes even in the off peak-periods. These off-peak trips occur at very low toll rates, further encouraging users.

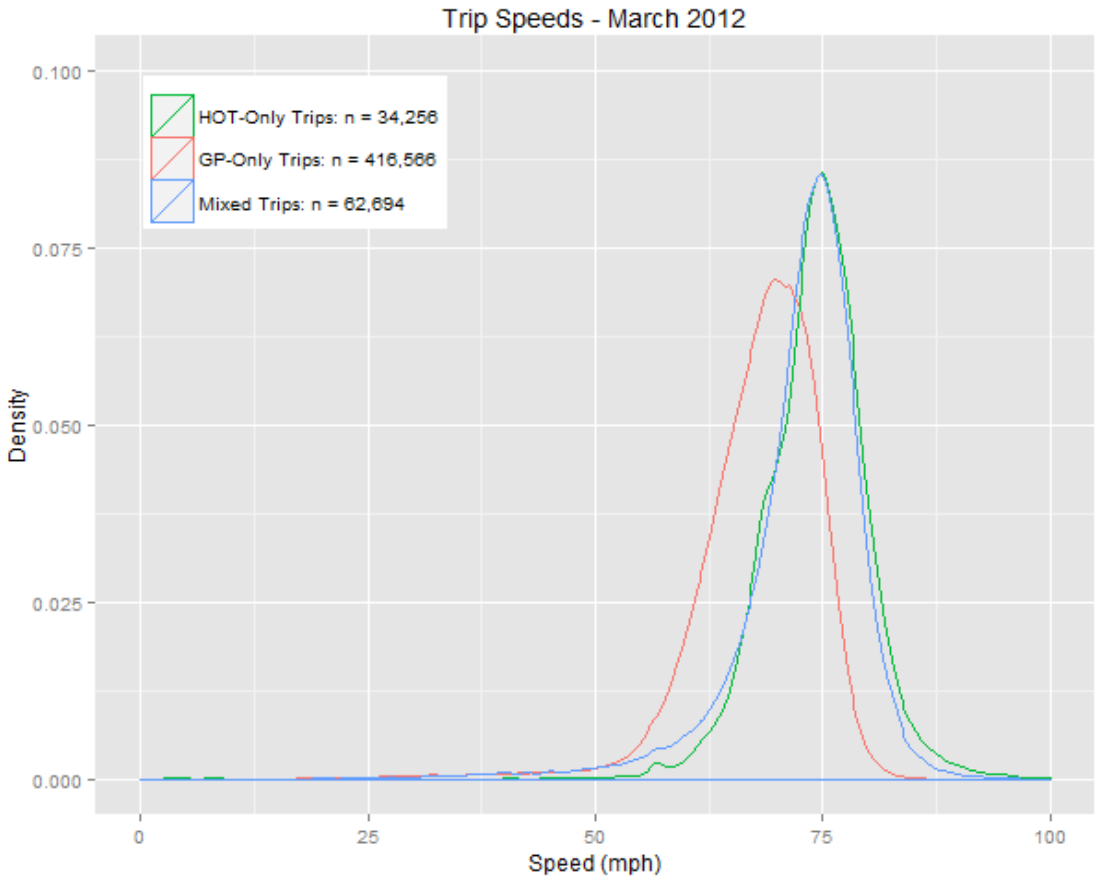


Figure 43: Speed Distribution for March 2012 Off-Peak Built Trips

Figure 44 illustrates the kernel density distribution of trip distances for all days and hours in the March 2012 data set. Note that the distance distributions include more distinct peaks as trip distance is a discrete measurement in this data. Because the vehicle detections occur only at the existing gantries, the potential distance measurements can only come from combinations of those gantries. The comparison is also not a direct comparison as the HOT lane detectors cover a longer length of the corridor than the GP lane detectors: the GP detectors extend across approximately 88% of the length of the HOT detector span.

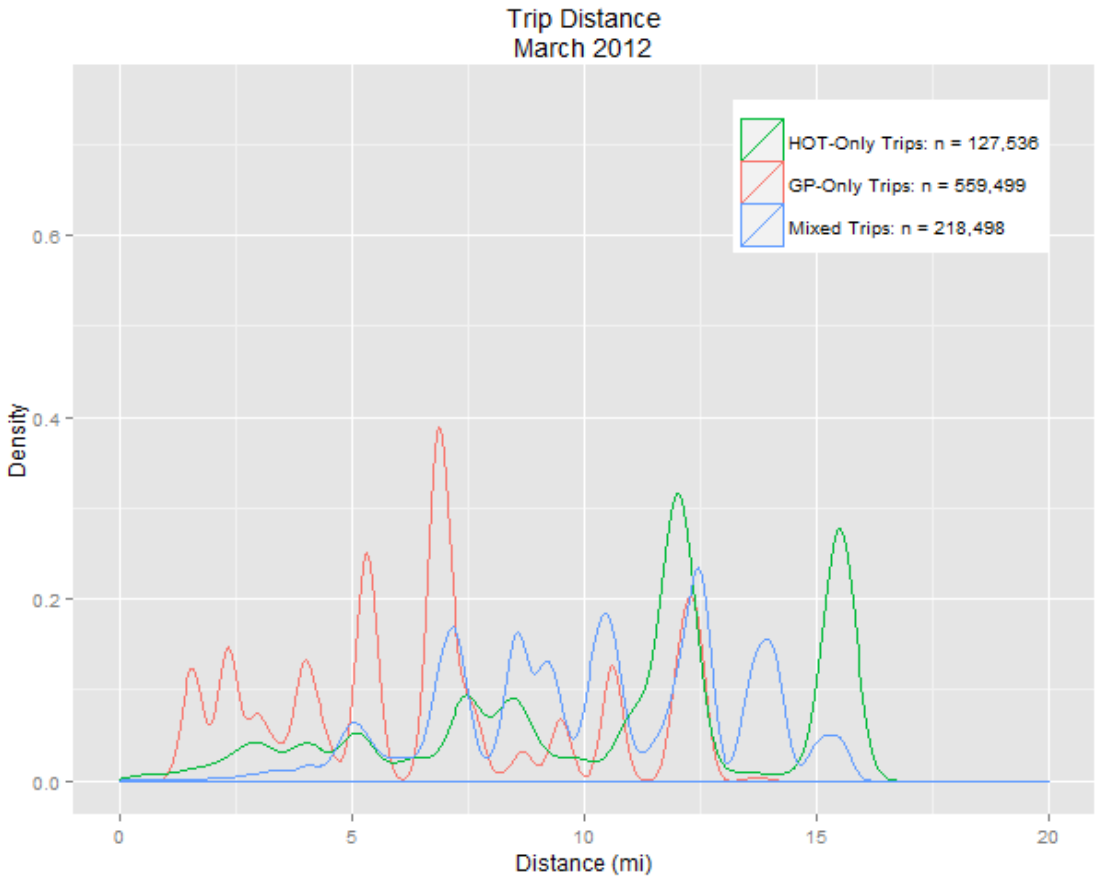


Figure 44: Trip Distance Distribution for March 2012

Figure 45, Figure 46, Figure 47, and Figure 48 illustrate the frequency of the different start and end locations for trips in the Express Lanes in March of 2012. Figure 45 shows the counts of trips that start and end at the six southbound Express Lane segments during the morning peak hours. A roughly equal number of trips start at Old Peachtree Road southbound and State Route 316 West. This is not surprising as these are the northernmost points of the facility. Combined, trips from the two entry locations make up a majority of all trips. The most frequent exit location is the I-285 segment, which is also understandable as it is the southernmost exit point. Entry and exit behavior in the GP lanes is different, as evidenced by Figure 46. Because there are no GP lane detectors on SR-316, researchers cannot see the breakdown of vehicles that enter from 316 versus those that enter from Old Peachtree Road. Here more trips begin at the second southbound GP gantry, in the Pleasant Hill Road section. Similarly, the largest number of trips end at the fifth southbound detector, before the exit to I-285. Unlike the HOT lane morning trips, which more often start and end at the extremes of the facility, the GP lane morning trips more frequently start and end within the facility.

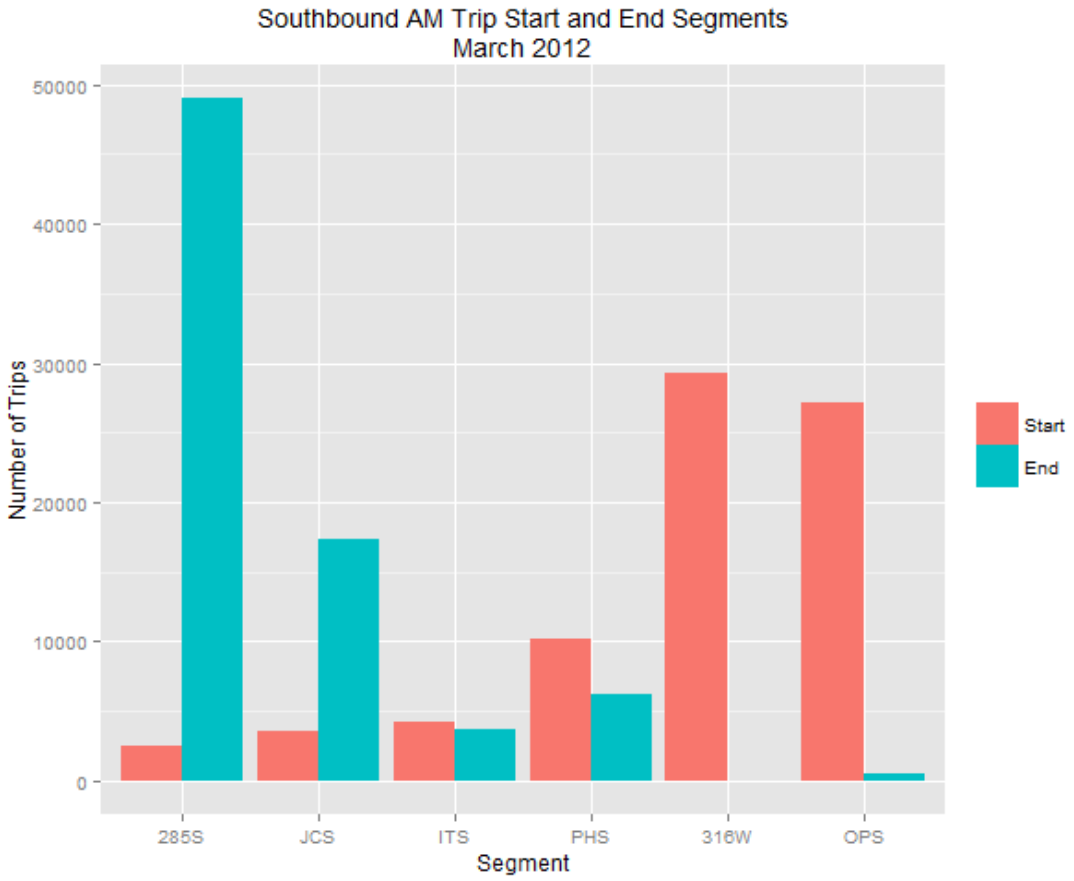


Figure 45: Start and End Segments for March 2012 AM Peak HOT Trips

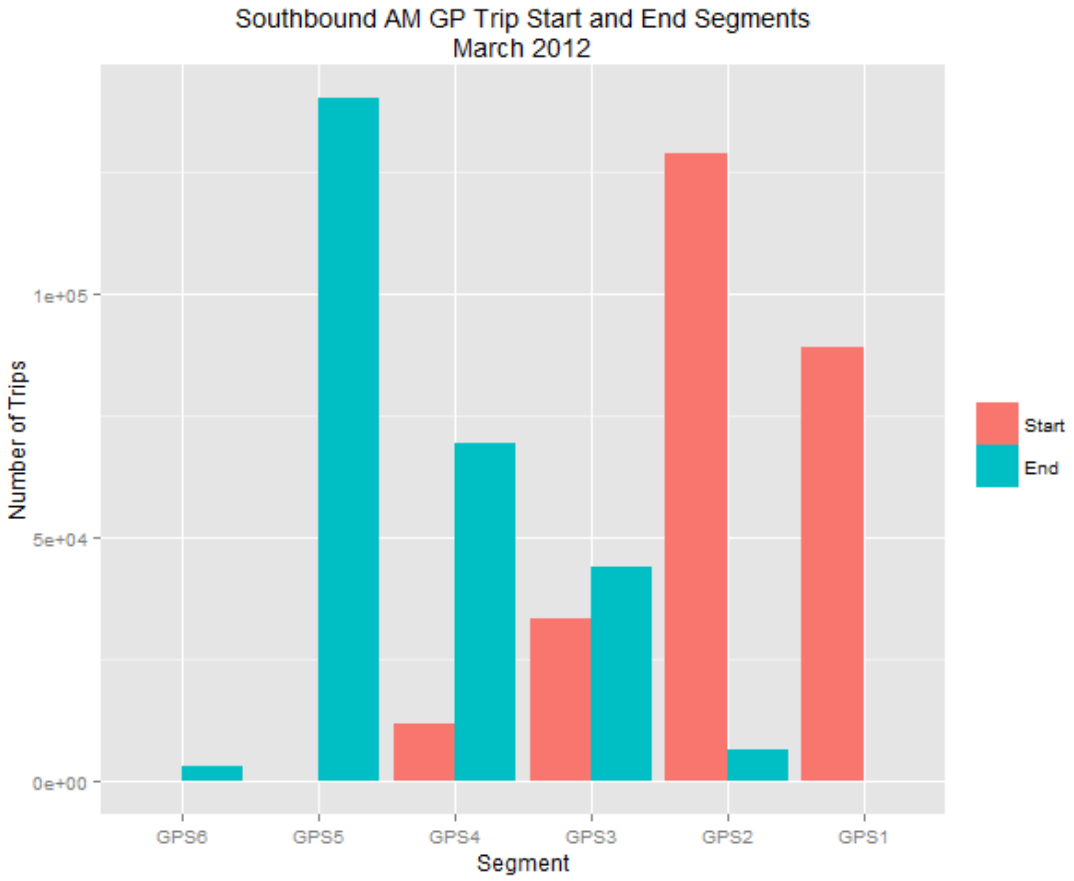


Figure 46: Start and End Segments for March 2012 AM Peak GP Trips

Figure 47 and Figure 48 present the frequency of start and end locations for HOT and GP lane northbound trips in the afternoon peak. The behavior here is different: the vast majority of HOT trips begin at a single location: I-285 northbound. This is to be expected as the facility does not fork at the southern end as it does at the northern end. SR-316 and Old Peachtree Road see the most trip exits, with more trips ending at Old Peachtree in this month. Figure 48 shows that the behavior in the GP lanes differs even more, in that roughly the same number of trips start at the first and second northbound gantries. The first gantry is south of the I-285 interchange, while the second is north of it. Most of the trip exits occur not at the last gantry or even the second-to-last gantry (both in the Old Peachtree Road section) but rather at the fourth northbound gantry, located in the Indian Trail segment of the facility. The sixth (Old Peachtree) gantry sees a similar, but slightly lower, number of trip end points. The counts and charts illustrate differences between the typical start and end points of HOT and GP trips, as well as between morning and afternoon peak period trips.

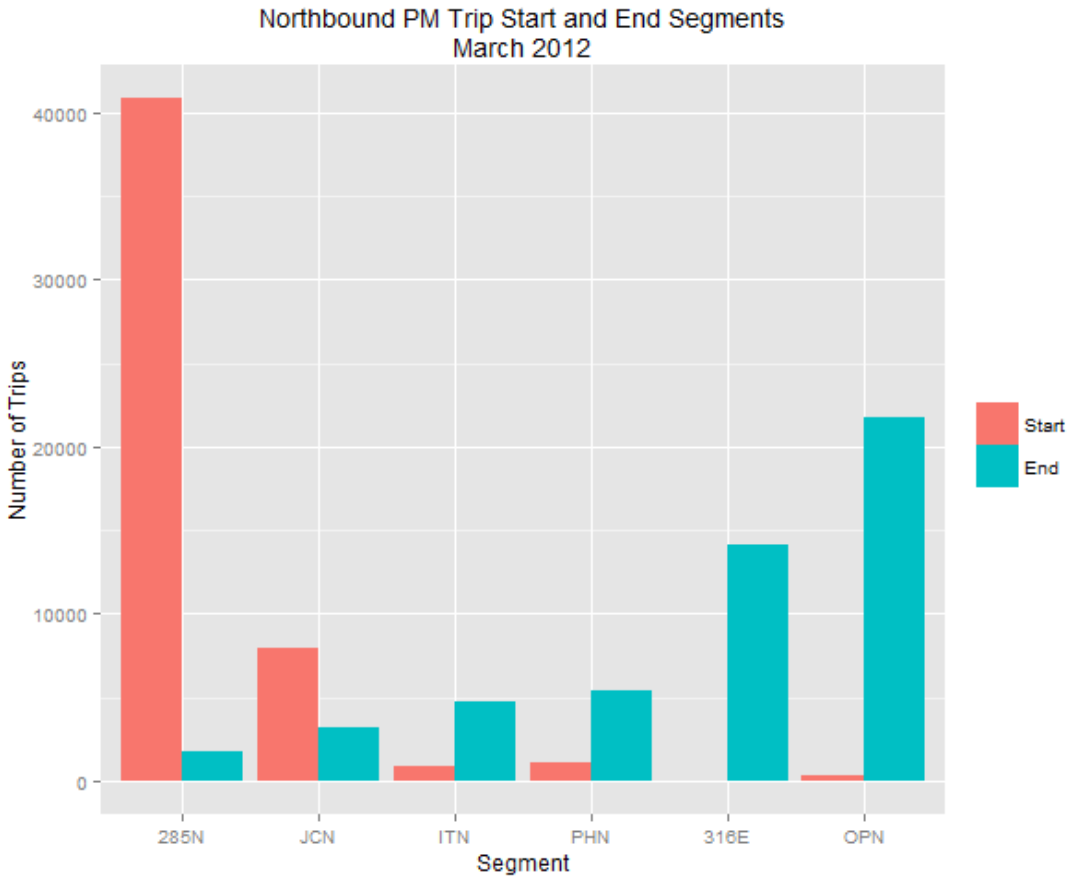


Figure 47: Start and End Segments for March 2012 PM Peak HOT Trips

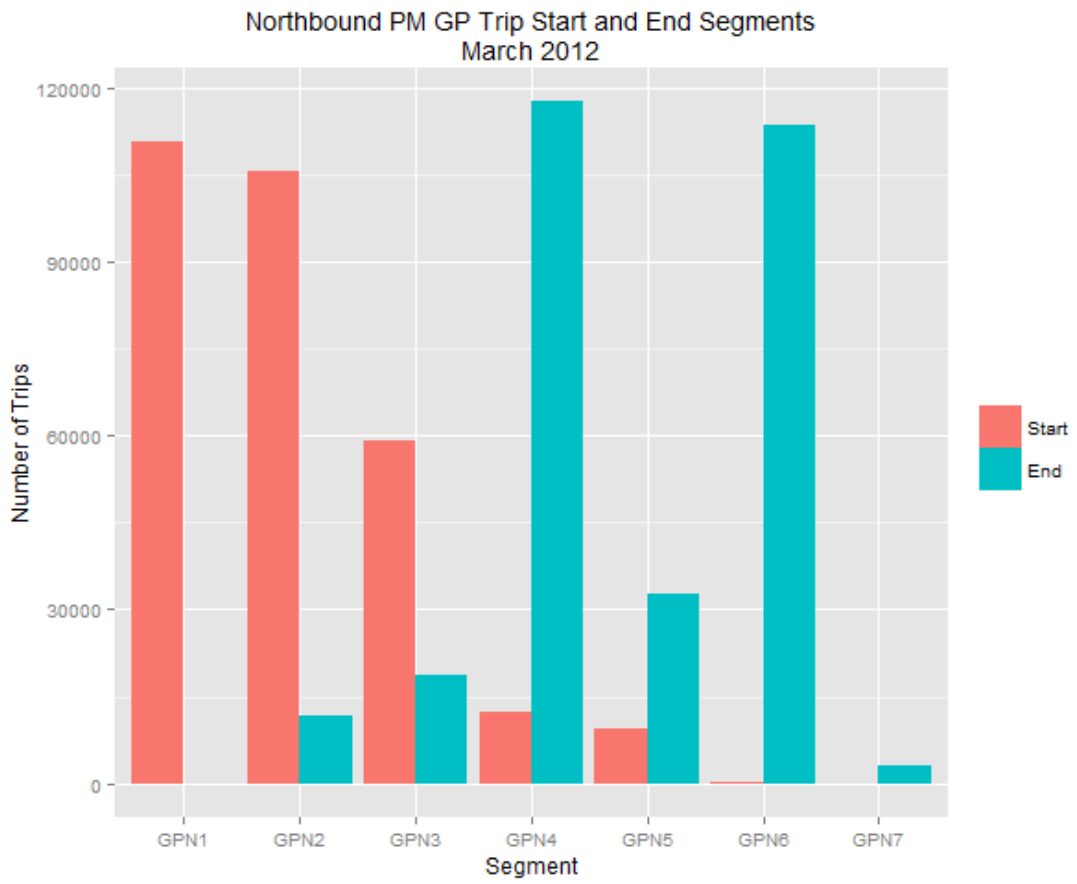


Figure 48: Start and End Segments for March 2012 PM Peak GP Trips

Travel Time Averages

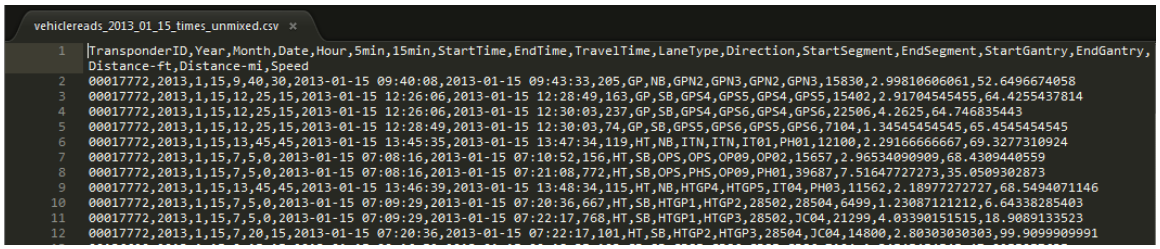
In addition to building trips from the disaggregate vehicle detections, these data were also used to find travel times of vehicles on the I-85 Express Lanes corridor. The algorithms that calculate vehicle travel times examine a single day's worth of detection data at a time. Vehicle detections are grouped by the unique identifier assigned to each Peach Pass transponder. The travel time script iterates through each detection for each transponder and identifies the gantry at which the vehicle was detected. Once all of the gantries have been identified, the script cycles through the possible combinations of gantries that constitute a trip. For those combinations that correspond to existing

detections, a travel time is calculated and stored. For example, consider a vehicle X that is detected at General Purpose northbound gantries 1, 3, and 6. The travel time script will cycle through the possible General Purpose northbound gantry combinations, such as gantry 1 to gantry 2, gantry 2 to gantry 3, gantry 1 to gantry 3, and so on. In this example, the detections of vehicle X would yield travel times from gantry 1 to gantry 3, gantry 1 to gantry 6, and gantry 3 to gantry 6. This method is used for all gantries in the HOT and GP lanes in both the northbound and southbound directions. The script does not calculate travel times between lane types; that is, if vehicle X is detected by an HOT lane gantry and then by a GP lane gantry, the software will not report a travel time between those two gantries. The results provide travel times for HOT-exclusive and GP-exclusive traverses by section. This is distinct from individual trip travel times, which measure the entire duration of the trip and may occur over both lane types.

Before outputting the travel time data, the travel time script applies filters to remove impossible results. The first of these verifies that the downstream detection occurs later than the upstream detection, thus avoiding "negative" travel times. The second filter removes travel times that are longer than an hour; the purpose of this filter being to remove trips that involve vehicles leaving and then returning to the freeway. The third filter verifies that both detections occur within the same twelve-hour timeframe (before noon or after noon). Finally, a function that detects "mixed" trips filters out travel times that involve detections in both lane types. In the case of vehicle X traveling between GP gantries 1 and 6, this last filter will check whether the vehicle was detected in the Express Lanes between those two gantries. If vehicle X spent some portion of that

trip in the HOT lane, that travel time is excluded from the final output. This is to ensure that the calculated travel times reflect the conditions of a single lane type.

The resulting daily travel time files give the unique transponder identifier, the date and time of the first and last detections, the travel time in seconds between those detections, the lane type and direction that the vehicle traveled in, the start and end gantries and the roadway segments those gantries correspond to, the distance in feet and miles, and the speed of the vehicle. Figure 49 below shows an example of a travel time output file.


vehiclereads_2013_01_15_times_unmixed.csv

```
TransponderID,Year,Month,Date,Hour,5min,15min,StartTime,EndTime,TravelTime,LaneType,Direction,StartSegment,EndSegment,StartGantry,EndGantry,
Distance-ft,Distance-mi,Speed
00017772,2013,1,15,9,40,30,2013-01-15 09:40:08,2013-01-15 09:43:33,205,GP,NB,GPN2,GPN3,GPN2,GPN3,15830,2.99810606061,52.6496674058
00017772,2013,1,15,12,25,15,2013-01-15 12:26:06,2013-01-15 12:28:49,163,GP,SB,GPS4,GPS5,GPS4,GPS5,15402,2.91704545455,64.4255437814
00017772,2013,1,15,12,25,15,2013-01-15 12:26:06,2013-01-15 12:30:03,237,GP,SB,GPS4,GPS6,GPS4,GPS6,22506,4.2625,64.746835443
00017772,2013,1,15,12,25,15,2013-01-15 12:28:49,2013-01-15 12:30:03,74,GP,SB,GPS5,GPS6,GPS5,GPS6,7104,1.34545454545,65.4545454545
00017772,2013,1,15,13,45,45,2013-01-15 13:45:35,2013-01-15 13:47:34,119,HT,NB,ITN,ITN,IT01,PH01,12100,2.29166666667,69.3277310924
00017772,2013,1,15,7,5,0,2013-01-15 07:08:16,2013-01-15 07:10:52,156,HT,SB,OPS,OPS,OP09,OP02,15657,2.96534090909,68.4309440559
00017772,2013,1,15,7,5,0,2013-01-15 07:08:16,2013-01-15 07:21:08,772,HT,SB,OPS,PHS,OP09,PH01,39687,7.51647727273,35.0509302873
00017772,2013,1,15,13,45,45,2013-01-15 13:46:39,2013-01-15 13:48:34,115,HT,NB,HTGP4,HTGP5,IT04,PH03,11562,2.18977272727,68.5494071146
00017772,2013,1,15,7,5,0,2013-01-15 07:09:29,2013-01-15 07:20:36,667,HT,SB,HTGP1,HTGP2,28502,28504,6499,1.23087121212,6.64338285403
00017772,2013,1,15,7,5,0,2013-01-15 07:09:29,2013-01-15 07:22:17,768,HT,SB,HTGP1,HTGP3,28502,JC04,21299,4.03390151515,18.9089133523
00017772,2013,1,15,7,20,15,2013-01-15 07:20:36,2013-01-15 07:22:17,101,HT,SB,HTGP2,HTGP3,28504,JC04,14800,2.80303030303,99.9099909991
```

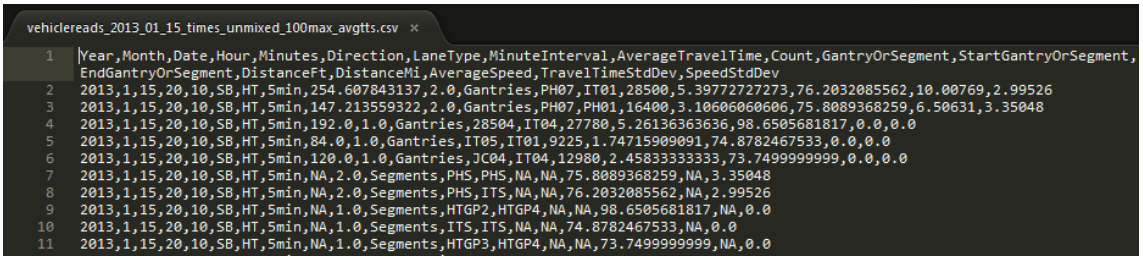
Figure 49: Sample Travel Time Output

After calculating travel times for each gantry combination in both the HOT and GP lanes, another script averages the travel times to provide an overview of traffic conditions along the corridor. This script reads one day's worth of travel time results at a time and provides average travel times, average speeds, and standard deviations of both measures between all of the various HOT and GP gantries.

Average speeds are calculated by summing up the total distance traveled by all vehicles between two gantries and dividing that by the total time taken by those vehicles. Average travel times are calculated using the harmonic mean method. The results are reported at the gantry level, for example Indian Trail gantry 1 to Pleasant Hill gantry 5, and also at the segment level: Indian Trail northbound to Pleasant Hill northbound. At the segment level, distances traveled by different vehicles may vary, and so the script

does not report travel times but only speeds. Each record also reports the number of travel times used in the calculation. The measures are calculated in bins of five minutes and fifteen minutes.

Figure 50 shows an example of the average travel time output in its raw form. This includes the date and time of the average travel time, along with an indicator of whether the result is calculated for a five-minute bin or a fifteen-minute bin. The output file also identifies whether the measure was at the gantry level or at the segment level. A segment is made up of multiple gantries, so the results from the segment level include the relevant results from the gantry level. The gantry level looks at all combinations of the individual vehicle detectors and provides averages of the travel times and speeds that occur between them. At the segment level, travel speeds between the gantries encapsulated within a given segment are averaged. Travel times are not reported at the segment level because the distances between the gantries within a segment varies. Gantry 1 and Gantry 3 may be farther apart than Gantry 2 and Gantry 4, for example, though all four are within Segment A.


vehiclereads_2013_01_15_times_unmixed_100max_avgtts.csv

```
1  Year,Month,Date,Hour,Minutes,Direction,LaneType,MinuteInterval,AverageTravelTime,Count,GantryOrSegment,StartGantryOrSegment,
   EndGantryOrSegment,DistanceFt,DistanceMi,AverageSpeed,TravelTimeStdDev,SpeedStdDev
2  2013,1,15,20,10,SB,HT,5min,254.607843137,2.0,Gantries,PH07,IT01,28500,5.39772727273,76.2032085562,10.00769,2.99526
3  2013,1,15,20,10,SB,HT,5min,147.213559322,2.0,Gantries,PH07,PH01,16400,3.10606060606,75.8089368259,6.50631,3.35048
4  2013,1,15,20,10,SB,HT,5min,192.0,1.0,Gantries,28504,IT04,27780,5.26136363636,98.6505681817,0.0,0.0
5  2013,1,15,20,10,SB,HT,5min,84.0,1.0,Gantries,IT05,IT01,9225,1.74715909091,74.8782467533,0.0,0.0
6  2013,1,15,20,10,SB,HT,5min,120.0,1.0,Gantries,JC04,IT04,12980,2.45833333333,73.7499999999,0.0,0.0
7  2013,1,15,20,10,SB,HT,5min,NA,2.0,Segments,PHS,PHS,NA,NA,75.8089368259,NA,3.35048
8  2013,1,15,20,10,SB,HT,5min,NA,2.0,Segments,PHS,ITS,NA,NA,76.2032085562,NA,2.99526
9  2013,1,15,20,10,SB,HT,5min,NA,1.0,Segments,HTGP2,HTGP4,NA,NA,98.6505681817,NA,0.0
10 2013,1,15,20,10,SB,HT,5min,NA,1.0,Segments,ITS,ITS,NA,NA,74.8782467533,NA,0.0
11 2013,1,15,20,10,SB,HT,5min,NA,1.0,Segments,HTGP3,HTGP4,NA,NA,73.7499999999,NA,0.0
```

Figure 50: Sample Travel Time Average Output

Figure 51 illustrates one month's worth of average travel times (in 15-minute bins) during the southbound morning peak for the entire corridor, from Old Peachtree Road to I-285. As these are travel times, not speeds, the endpoints were the first and last gantries on the HOT lane. The results are presented in fifteen minute bins for each

weekday of January, 2012. Mean travel times in the figure are consistent until 7:00AM; at that point, there is greater variation in the daily travel time averages. This variation lasts until roughly 8:45AM, at which point the daily average corridor travel times are consistent again. Figure 52 illustrates the corresponding travel times for the northbound afternoon peak, from the first gantry in the I-285 segment to the final gantry in the Old Peachtree Road segment. The results are much more consistent, with fewer 'slow' days in which the average travel time was outside of a narrow range.

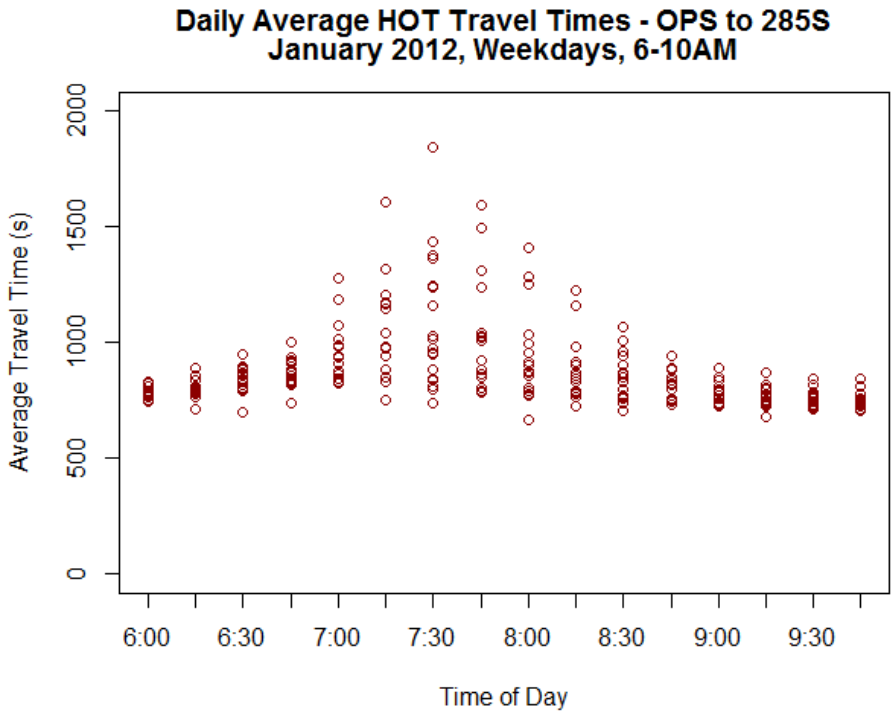


Figure 51: Daily Average HOT Travel Times - Southbound

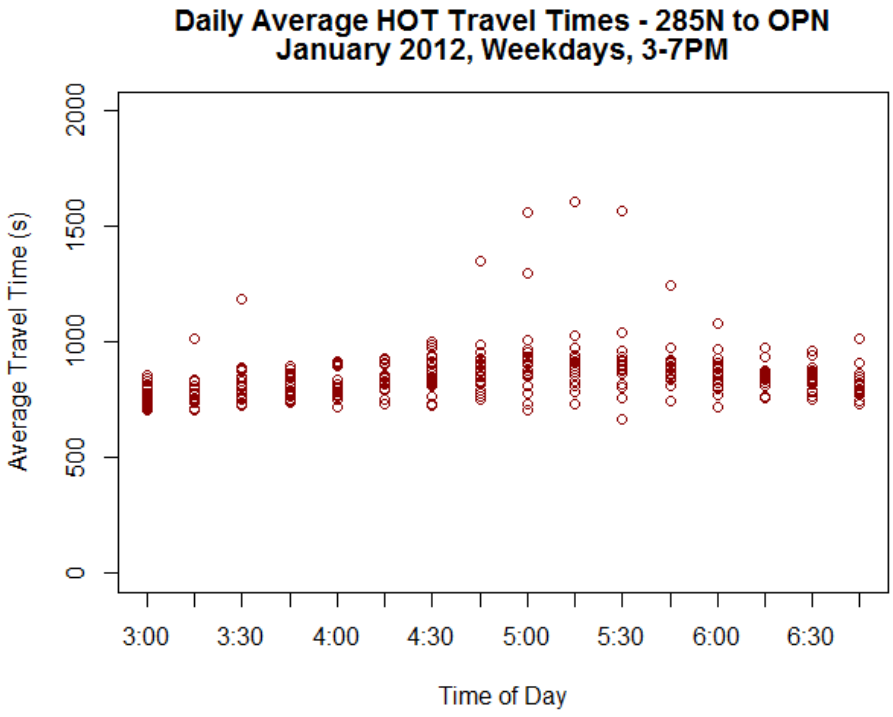


Figure 52: Daily Average HOT Travel Times - Northbound

Figure 53 and Figure 54 present the average daily travel times on the General Purpose lanes for the southbound morning and northbound afternoon peaks, respectively. Note that the maximum value on the y-axis for these figures is twice that of the two previous figures. Here, the relevant detectors are the first and last GP scanners in each direction. The results are much more varied, with no 'narrow' consistent interval in either timeframe. This may be an artifact of the grouping together of all GP lanes for the travel time calculations, rather than examining each GP lane individually. The northbound plot shows slightly tighter clustering at the beginning of the study period, until 4:00PM, but both charts show little consistency otherwise.

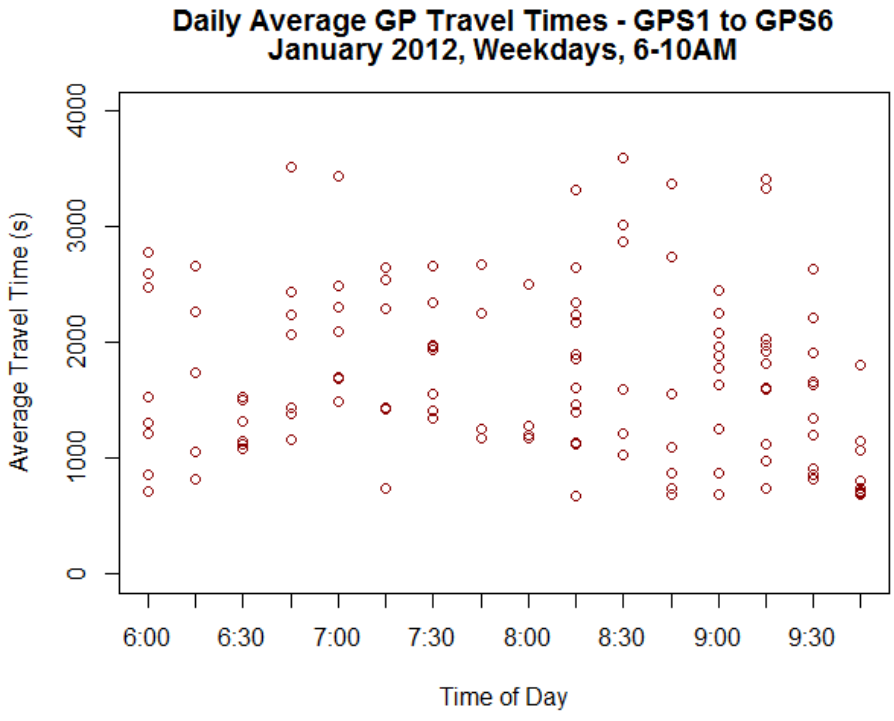


Figure 53: Daily Average GP Travel Times - Southbound

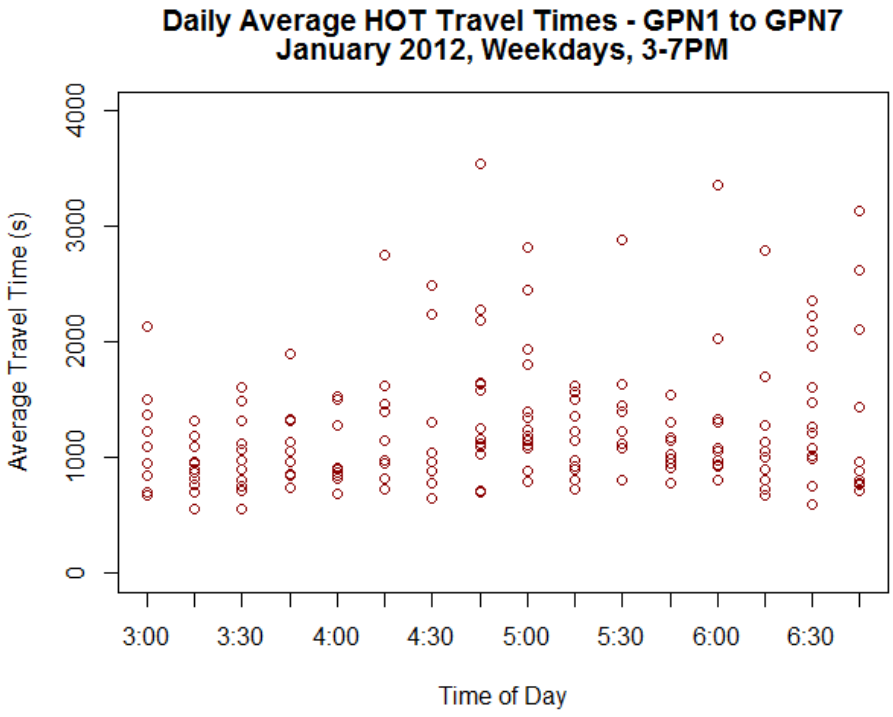


Figure 54: Daily Average GP Travel Times - Northbound

Unique Transponder Counts

Another script examines the disaggregate vehicle detection data to count the number of unique Peach Pass transponders in the corridor. The script groups the results by the different entry and exit combinations and reports counts in five or fifteen minute intervals. The script reads an individual day's vehicle detections and finds the lane and corridor segment in which each detection occurred. The various possible entry and exit combinations for the corridor record a detection for that transponder in that five or fifteen minute interval. For example, if Vehicle X is detected in the HOT lanes in the Jimmy Carter Boulevard segment, the script would note the detection for the Jimmy Carter segment, the I-285 to Jimmy Carter segment, the I-285 to Indian Trail segment, the I-285 to Pleasant Hill segment, and so on. Similarly, it would note the detection for the Jimmy Carter to Indian Trail segment, the Jimmy Carter to Pleasant Hill segment, and the Jimmy Carter to Old Peachtree segment. The script proceeds in this manner, noting detections for all of the segment combinations that include the location of the detection.

Once the script has finished reading the detection file and has identified all of the transponders and the segments in which they were found, the script cycles through the results and prints out the number of unique transponders found along with the date and time, the time interval (five or fifteen minutes), and the start and end segment name for each potential entry and exit combination. Figure 55 below shows the first few lines of a transponder count output file. In this case, lines 2 through 7 show counts for segments beginning with the JC01 gantry (Jimmy Carter gantry number 01) and ending at various gantries within the 285 segment. As the segment under examination increases in length, more vehicles are counted within the five minute interval. The segment from JC01 to

28507, the shortest distance, contains 11 unique Peach Pass transponders, while the longest segment, from JC01 to 28502, contains 20.

vehiclereads_2013_01_15_transcounts_peak.csv

```
1  Year,Month,Date,Hour,minutebin,minutes,direction,
   StartGantrySegment,EndGantrySegment,Count
2  2013,1,15,6,5min,0,SB,JC01,28502,20
3  2013,1,15,6,5min,0,SB,JC01,28503,16
4  2013,1,15,6,5min,0,SB,JC01,28506,12
5  2013,1,15,6,5min,0,SB,JC01,28507,11
6  2013,1,15,6,5min,0,SB,JC01,28504,15
7  2013,1,15,6,5min,0,SB,JC01,28505,13
8  2013,1,15,6,5min,0,SB,JC03,JC01,15
9  2013,1,15,6,5min,0,SB,JC03,JC02,15
10 2013,1,15,6,5min,0,SB,JC03,28502,29
11 2013,1,15,6,5min,0,SB,JC03,28503,25
12 2013,1,15,6,5min,0,SB,JC03,28506,21
13 2013,1,15,6,5min,0,SB,JC03,28507,20
14 2013,1,15,6,5min,0,SB,JC03,28504,24
15 2013,1,15,6,5min,0,SB,JC03,28505,22
16 2013,1,15,6,5min,0,SB,JC02,28502,20
17 2013,1,15,6,5min,0,SB,JC02,28503,16
```

Figure 55: Sample Transponder Count Output

Figure 56 shows the transponder count results for the southbound morning peak period, from Old Peachtree Road to I-285. Here the number of vehicles detected in the facility in fifteen-minute intervals increases consistently until the peak-of-the-peak at 7:30AM. As the detection counts decrease afterwards, the counts become less consistent. Figure 57 illustrates the transponder count results for the northbound afternoon peak. Again, a steady increase in vehicle counts on the corridor can be seen until the peak-of-the-peak at 5:30PM. Here the counts are more consistent overall, mirroring the average travel time plots presented previously. Note that the two days that yield low transponder counts are January 2 (the day after New Year's Day) and January 16, Martin Luther King, Jr. Day (a federal holiday).

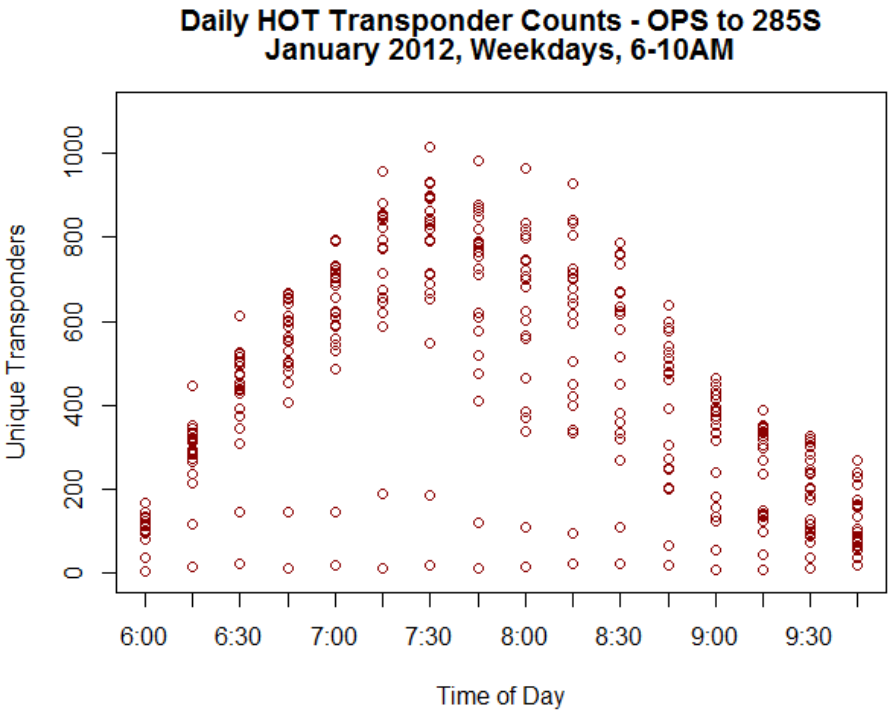


Figure 56: Daily HOT Transponder Counts - Southbound

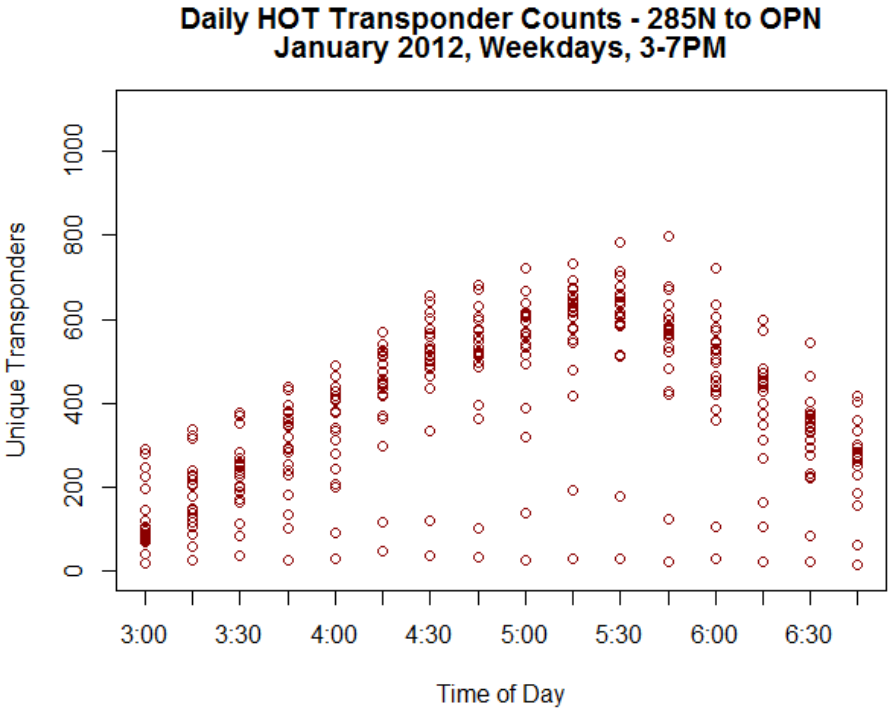


Figure 57: Daily HOT Transponder Counts - Northbound

Figure 58 and Figure 59 present the transponder counts in the GP lanes for the southbound morning and northbound afternoon peak periods, respectively. Unlike the average travel time plots for the General Purpose lanes, the transponder counts show more consistency throughout the fifteen-minute intervals represented here. The peaks-of-the-peaks, represented by the intervals with the highest transponder counts, appears to match those of the HOT lanes. In both lane types, this peak occurs around the 7:30AM interval for the southbound direction and the 5:30PM interval for the northbound direction.

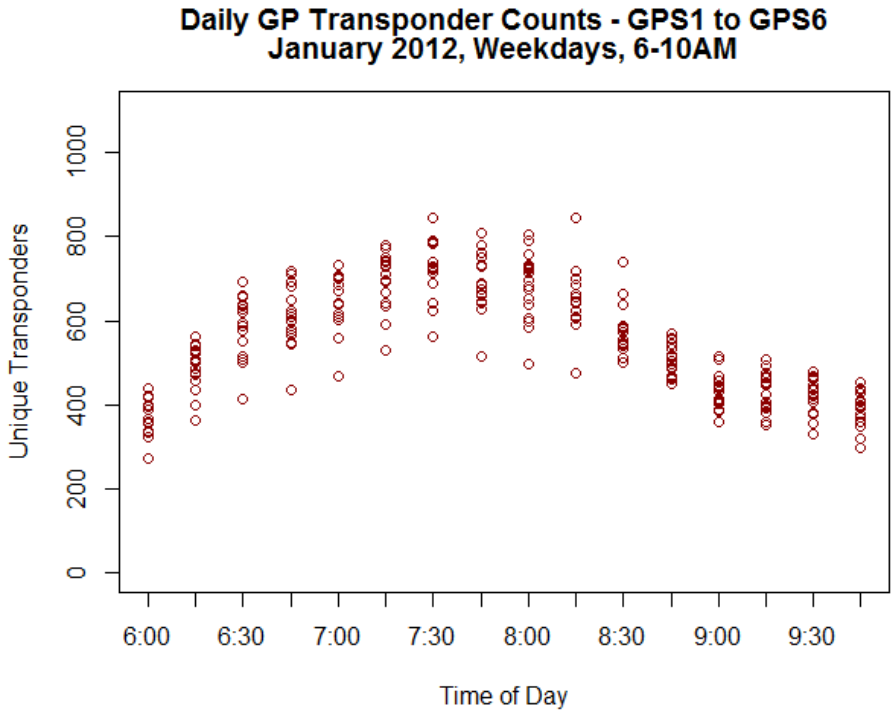


Figure 58: Daily GP Transponder Counts - Southbound

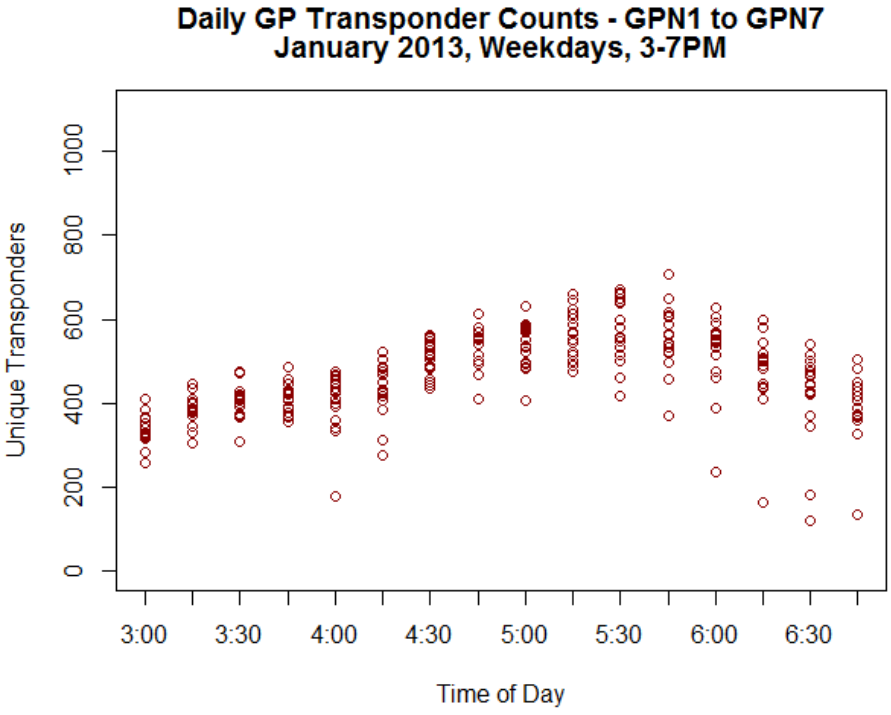


Figure 59: Daily GP Transponder Counts - Northbound

Construction of the Analytical Data Set

After using the disaggregated detections to construct vehicle trips, calculate travel time averages, and count distinct transponder reads, the suite of scripts then joins these results to data from other SRTA-provided streams. This section provides an overview of the data streams and the joining processes.

The purpose of the processing of the various data streams described in this chapter is to generate a comprehensive file for use in the subsequent analytical work. This will allow for the proposed investigation into user behavior and decision making at the trip level which incorporates the various data streams provided by SRTA and Epsilon. This file will include one observation per vehicle trip on the corridor. The structure of each observation is built around the proposed dependent variable: the use of the Express Lanes on a given trip. Additional data elements will indicate whether the trip was exclusively in the toll or general purpose lanes, or whether it involved a combination of both lane types (resulting in a partial Express Lane trip). Each observation will also include a set of independent variables involving trip characteristics, corridor conditions, and household demographic data. The trip characteristics include the time of the trip, the origin and destination along the corridor, and the Express Lane toll rate at the time of the trip. The corridor condition variables include the average speeds and transponder counts in both lane types. The demographic data include household income, size, education level, and head of household age. Additional variables that fall into these three categories, and the process by which they are joined in the data set, are discussed in the following paragraphs.

The building of the analytical file that incorporates the different data streams begins with the set of vehicle trips that were constructed from the individual vehicle detections. The first step is joining these constructed trips to the Express Lane Trip stream summary data provided by SRTA. This Trip stream data provides one record per Express Lane trip, and includes the start and end points, the entry and exit times, the toll mode and the amount paid, the transponder identifier, the vehicle plate number, and more. Of these fields, the toll amount and toll mode are brought into the analytical dataset. This allows researchers to identify the trips that were taken in carpool mode (toll mode = 'NON-TOLL') and thus were charged no toll. Similarly, it provides the actual amount of toll paid by the toll-mode users. The constructed trips are joined to the SRTA Trip data by the trip date, the start time, and the transponder identifier. The trip date and transponder identifier must be the same, while the start time must be within the same five-minute interval. This is to allow for minor differences between the two trip data sets: the constructed trips may have an extra or a missing gantry read, for example, that would change the start time slightly.

After the Trip stream join, the joining script then brings in the Toll Rate data provided by SRTA. This allows for Express Lane toll rates to be reported for trips that were taken in the General Purpose lanes and so did not have records in the Trip summary data. This is how researchers are able to find the toll a user would have paid had they used the HOT facility instead of the unpriced lanes. This join also reports the maximum daily toll, allowing researchers to identify Express Lane trips that were taken at the maximum toll rate. These trips are important to identify because the maximum toll rate charged on the corridor was limited at the opening and has been increasing gradually

since; this was previously addressed in the Data Sources chapter. This join also relies on the trip date and the five-minute interval of the start time, as well as the section of the corridor on which the trip occurred, because toll rates differ by corridor section.

The next join brings the average travel time and speed data into the dataset. As discussed earlier, these records are harmonic means of the travel times between all of the entry and exit combinations on the Express and General Purpose lanes. This allows researchers to compare average travel times and speeds across the different lane types for all of the dates and times for which data are available. This dataset also includes standard deviations of the speeds and travel times. These results are joined on date, time, fifteen minute interval, and start and end points. The transponder count data is joined in the same manor; these data are also reported for all start and end point combinations for both the Express and GP lanes.

The next step is to join the constructed trips with account data and status. The script locates the account data files associated with the date of the trip; if the account data is missing for that day (due to a corrupted or missing data file), the script identifies the first valid account file for that month. The script reads in the account and transponder files for the purpose of identifying the account type: personal or corporate. From the transponder ID associated with a trip, the script identifies the account and returns the account type.

The final function in the joining script cleans up entries that do not have complete entries for all of the data fields. If the script could not successfully create a join to the toll rates, average travel times, or travel counts, the row is removed from the dataset. These records have no identified toll rate for the time and location, no identified maximum toll

rate, average speeds of 0 mph, or no transponders identified at that time on the corridor. The narrowing of trip data that results from this process is explored in Chapter [7], under the Data Pairing and Loss heading. Finally, a fundamental step in the joining process to create this analytical file is matching the SRTA records with the marketing demographic data based on household locations from the vehicle registration database. This step is very complex, and is therefore described in more detail in the following chapter: Connecting SRTA Data to Epsilon Data.

CHAPTER 5

CONNECTING SRTA DATA TO EPSILON DATA

To bring demographic data into the working data set, the marketing data must be joined to the trips constructed from the SRTA lane use data. This multi-stage process is outlined in this chapter. The chapter begins by discussing the registration database matching portion of the process, which is performed by the Georgia Tech Research Institute. The next section outlines the steps required to match those results with the Epsilon demographic data. A summary of the results of the pairing process follows, along with a discussion of the commutershed restriction employed in this research. The chapter then provides an overview of selected demographic characteristics of the paired data set. Finally, the last section compares the paired households with the full Epsilon marketing data set to identify any potential differences.

GTRI Vehicle Registration Database Pairing

In April of 2014, all of the registered Georgia license plates with Peach Pass accounts were provided to the Georgia Tech Research Institute (GTRI) for matching against the state vehicle registration database. The data were matched with a blind process that created a link between observed plates and the privately sourced data without explicitly connecting license plates with registration database data. This process involved 983,860 unique plates sent to GTRI, sourced from both the Trip data stream and the Account data stream. The structure of the Account data stream prevented joining transponders to vehicles when an account had more than one of either; this issue is

further discussed in the Data Quality and Treatment chapter of this dissertation. A total of 521,159 license plates, from accounts with only one transponder and one vehicle, were selected for registration database pairing from the Account stream. At the time, it was understood that the Trip data stream did not have this issue; a trip record includes both the unique transponder identifier and the vehicle plate number and so researchers believed that these pairings were unique. As a result, 689,692 license plate and transponder combinations were identified from the Trip data stream. The final figure of 983,860 license plates included substantial overlap between the Account and Trip data sets. On May 23, 2014, GTRI delivered the set of household addresses based on the vehicle registration database. The set of addresses successfully matched and returned by GTRI included 518,099 non-unique records; that is, addresses appeared multiple times in the returned data set.

Epsilon Pairing Script Process

The script which pairs the SRTA lane use data with the Epsilon marketing data then reads the Account stream files for a specified date. Account data are received on a daily basis and can change just as frequently; accounts and transponders, for example, may be in 'active' status one day and then inactive the next. The script looks at those accounts and transponders that are active on a given day. For those accounts with one active transponder and one vehicle, the transponder ID and vehicle plate are paired and stored. Accounts with multiple vehicles and transponders are stored in a separate file. These accounts lack a one-to-one join between vehicles and transponders, and so they must be handled separately. Only accounts with one transponder and one vehicle or

accounts whose vehicles are all registered at the same address can successfully be paired with the Epsilon data.

The pairing scripts then examine the Trip summary stream that SRTA provides. This is the data set that lists all of the Express Lane trips, and provides transponder and license plate data for each observation. This stream was initially thought to be capable of addressing the many-to-many issues in the Account data; because each record has a single license plate and transponder, researchers thought that this would provide the one-to-one join that the Account data lacks. This turned out not to be the case; again, the Data Quality and Treatment chapter describes these issues in greater detail. The pairing scripts find only those Trip stream records for which one-to-one relationships between plates and transponders exist, and then adds these to the pool of unique pairings.

The next step in the matching process is reading the registration database records delivered by GTRI. As discussed in the previous section, researchers had sent GTRI a list of the license plates found in the SRTA data streams, along with an obscured key identification field for each record. The key field allowed GTRI to return the registration results without including the license plates. The script reads and temporarily stores the registration database records. Because of the multiple transponder and multiple vehicle issue discussed above, the GTRI matches are separated into those corresponding to single-vehicle, single-transponder accounts and those corresponding to many-to-many accounts.

After reading in the registration database matches, the script reads in the full set of marketing demographic data. These data include the addresses for each household. The script then iterates through both datasets to find those records with matching address

data. Researchers put the addresses through a standardization process with the goal of increasing the rate of 1:1 matches. This involved changing 'Rd' to 'Road,' 'St' to 'Street,' and dozens of other changes. Researchers also examined the records for misspellings and other issues that may cause matches to fail. The total number of records that saw exact matches was 148,352; this represents 28.6% of the GTRI registration data set and 42.5% of the Epsilon household data set.

After pairing with the registration database records, the resulting data are filtered to include only those within the I-85 commutershed. This commutershed was defined by Khoeini (2014) in her dissertation and it identifies the ellipse in which 95% of the corridor users have registered their vehicles. The commutershed is outlined below in Figure 60. Out of a total of 518,099 registration database records that GTRI returned, 417,350 are located within the commutershed. With the I-85 commutershed restricting the set of addresses, the resulting match count between the GTRI registration database and the Epsilon household marketing data is 135,170 records. This represents 26.0% of the GTRI registration data set and 38.7% of the Epsilon household data set.

At this point, the GTRI registration database records have been paired with the Epsilon household demographic records. To successfully complete the pairing process, these results must also be paired with the SRTA plates and transponders that were identified at the start of this section. The script begins by pairing the one-to-one transponders and plates with the registration database results provided by GTRI. From here, those one-to-one records with GTRI data are then narrowed to records that also include successful Epsilon pairings. These one-to-one records are then supplemented by accounts with multiple transponders that are ultimately associated with a single address.

The scripts then apply the commutershed filter to further narrow the result set. The next section provides an overview of the results of this pairing process at both the unrestricted and commutershed-restricted levels.

Results of SRTA-Epsilon Pairing Process

Table 18 presents a snapshot of the pairing results from a single day of account data: January 1, 2014. The table presents the rates of matching which occur between the SRTA transponder records, the GTRI registration database, and the Epsilon marketing data set. The full population of transponders under examination consists of all active transponders listed in the SRTA Account data stream; this population consists of 436,753 active transponders. Of those, nearly 64% originate from Peach Pass accounts that have a single transponder and a single vehicle. Thus for this sample of transponders, the pairing scripts can directly associate a single transponder with a single plate. Almost 85,000 of these one-to-one transponders come exclusively from the account data, in which the relevant accounts have a single associated vehicle and a single associated transponder. Nearly twice as many transponders originate from the Trip summary stream; these transponders were paired with a single plate within the SRTA data from the opening of the facility to the end of 2014. A subset of the transponders with one-to-one plate matches, consisting of 46,836 transponders, were common to both the Account and the Trip data streams. These transponders were counted only once in the final tally.

The next step in the pairing process matches the transponders with address records from the GTRI registration database. Of the 278,984 transponders with a one-to-one plate relationship, 254,280 can be paired with GTRI records. This set is supplemented by transponders that do not have a one-to-one plate relationship, but are

associated with a single GTRI address. These transponders have one-to-many, many-to-one, or many-to-many relationships with the license plates in their parent accounts. Once these license plates were paired with the GTRI address data, however, the resulting addresses were the same and so researchers could assign an address to these accounts and transponders. The share of transponders from these accounts is roughly half that of the one-to-one accounts, and they represent 29% of the overall transponder population.

The final step in the pairing process matches the GTRI-paired transponders with the Epsilon demographic data based on the addresses in the GTRI and Epsilon data sets. This step creates the largest loss of data: of the 380,976 transponders with GTRI matches, only 98,213 were successfully paired with Epsilon data. The majority of these, 55,686, come from the one-to-one transponder list. The analyses that follow will employ about data recorded from about 100,000 vehicles and household using the corridor, but this represents only about 23% of the transponders traversing the corridor. The following sections discuss the limited data availability and potential uncertainty issues.

Table 18: Snapshot of SRTA-Epsilon Matched Transponders

	Transponders	**% of Total**
All active transponders in SRTA Account data (1/1/2014)	**436,753**	**100.0%**
Active transponders with one-to-one plate matches	**278,984**	**63.9%**
Account stream-only transponders with one-to-one plate matches	84,658	19.4%
Trip stream-only transponders with one-to-one plate matches	147,490	33.8%
Transponders with one-to-one plate matches found in both streams	46,836	10.7%
Total GTRI HH paired transponder count from all sources	**380,976**	**87.2%**
One-to-one transponders paired to GTRI HH data	254,280	58.2%
Additional one-to-many, many-to-one, or many-to-many transponders paired to GTRI HH data	126,696	29.0%
Total Epsilon paired transponder count from all sources	**98,213**	**22.5%**
One-to-one transponders paired to Epsilon data	55,686	12.7%
Additional one-to-many, many-to-one, or many-to-many transponders paired to Epsilon data	42,527	9.7%

I-85 Commutershed Restriction

The registration-matched dataset provided by GTRI (518,169 records) was a much larger data set than the set of Epsilon marketing demographic records (349,134 records). This contributed to the low match rate between the Epsilon data and the GTRI registration address data: geographically, the GTRI data cover a larger area and includes many households outside of the I-85 commutershed. The Epsilon data purchase, on the other hand, was restricted to households within the I-85 commutershed. In an attempt to investigate the match rate of household records between the Epsilon and GTRI data, the author restricted the registration data to zip codes within the I-85 commutershed as

defined by Khoeini in her doctoral dissertation (2014). The ellipse was designed to capture 95% of the Express Lane users identified in the license plate collection study (Guensler, et al., 2013). Figure 60 illustrates the commutershed defined by Khoeini and the selection of the zip code regions that intersect it. The dark blue ellipse represents the Express Lane commutershed, while the lighter blue areas beneath the ellipse are the zip code regions that intersect with the commutershed.

The selection identified 132 zip code regions as intersecting with the commutershed. The author then restricted the GTRI registration database records to those from these 132 zip codes and re-matched the records with the Epsilon marketing data. The effect of this restriction on the GTRI and Epsilon match rates can be seen in Table 19 and the discussion below.

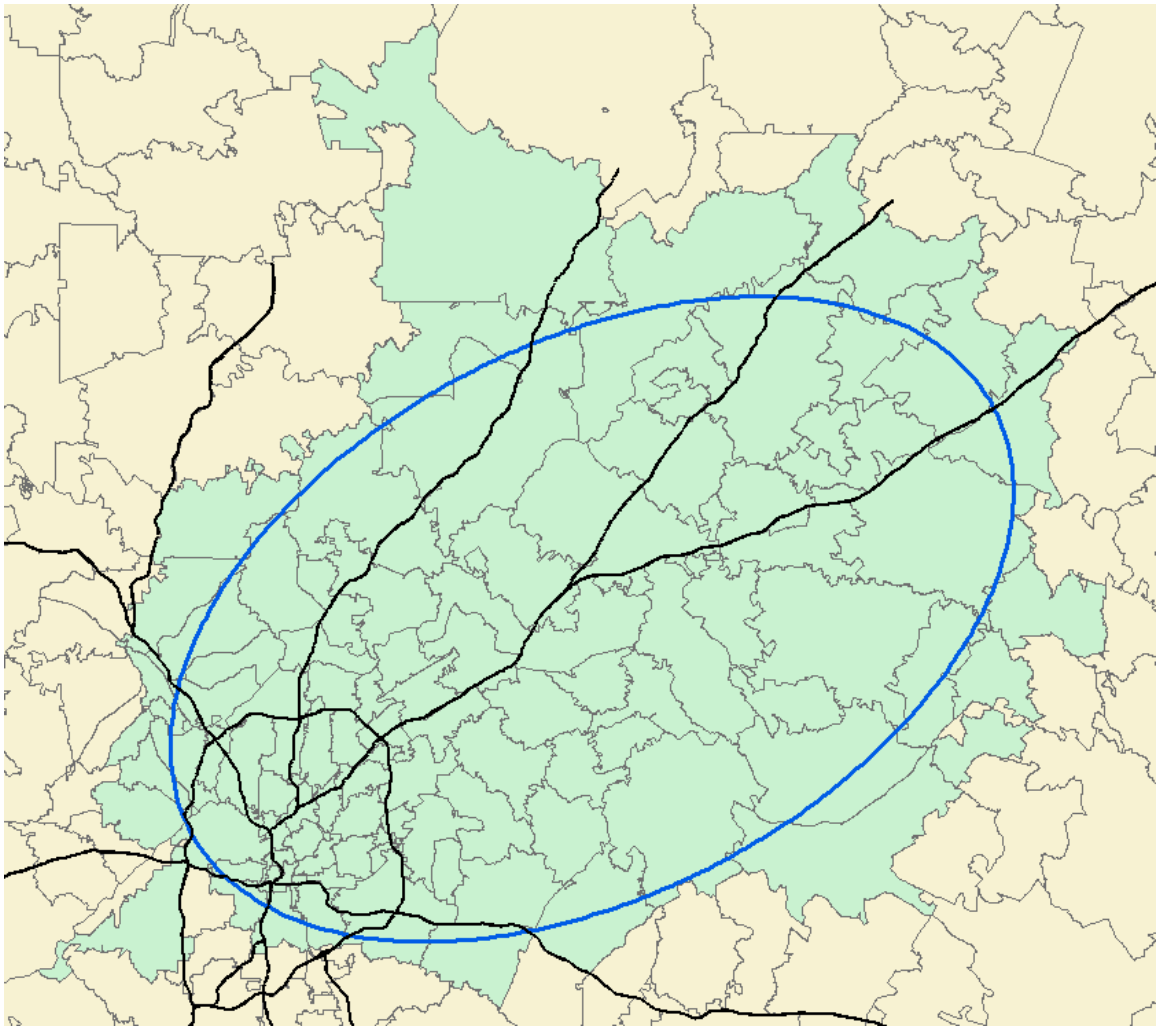


Figure 60: Zip Code Regions Intersecting I-85 HOT Commutershed

Table 19 provides a similar snapshot of SRTA-Epsilon pairing rates, this time restricted to transponders and households that fall within the I-85 commutershed. The differences begin at the GTRI-pairing step of the process; here, approximately 80,000 fewer transponders were successfully paired to the GTRI registration database results. That drop off rate is not nearly as high for the Epsilon pairing step; there 6,029 transponders were excluded by the commutershed restriction. This is to be expected as researchers specifically targeted the Epsilon data purchase to households within the I-85 commutershed. One motivation for restricting match results to the commutershed was

the hope that such a restriction would improve the match rate in the sample, and it is true that the the proportion of GTRI-matched transponders that can be paired with Epsilon data is higher within the geographically restricted sample. Of those GTRI-matched commutershed transponders, 33.11% were successfully paired with Epsilon records. Of the GTRI-matched transponders that have no commutershed restriction, 25.78% of the sample was successfully paired. The overall rate of transponder pairing among the entire active transponder population is similar with or without the commutershed restriction; that restriction removes 6,029 transponders from the final Epsilon-matched sample.

Table 19: Snapshot of SRTA-Epsilon Matched Transponders in Commutershed

	Transponders	**% of Total**
All active transponders in SRTA Account data (1/1/2014)	**436,753**	**100.0%**
Active transponders with one-to-one plate matches	**278,984**	**63.9%**
Account stream-only transponders with one-to-one plate matches	**84,658**	**19.4%**
Trip stream-only transponders with one-to-one plate matches	147,490	33.8%
Transponders with one-to-one plate matches found in both streams	46,836	10.7%
Total commutershed GTRI HH paired transponder count from all sources	**278,364**	**63.7%**
One-to-one transponders paired to commutershed GTRI HH data	149,053	34.1%
Additional one-to-many, many-to-one, or many-to-many transponders paired to commutershed GTRI HH data	129,311	29.6%
Total Epsilon paired commutershed transponder count from all sources	**92,184**	**21.1%**
One-to-one commutershed transponders paired to Epsilon data	52,406	12.0%
Additional one-to-many, many-to-one, or many-to-many commutershed transponders paired to Epsilon data	39,778	9.1%

Overview of Paired Households

This section presents select distributions of demographic characteristics for the paired households identified by the matching process. These measures include household income, household size, household education, head of household age, home ownership category, and dwelling unit type. This section also compares the results to Census Bureau estimates of household demographics in the City of Atlanta geography, taken from the five year American Community Survey results (2013).

The first data set characteristic examined was the number of active transponders associated with each household. Figure 61 below illustrates the distribution of transponders per household within the matched Epsilon dataset. The pairing date for this distribution was January 1, 2014; as mentioned above, SRTA to Epsilon pairing can vary day-to-day due to the changing nature of the Account data stream. A plurality of matched households, just over 40%, have one associated Peach Pass transponder. Few households, 11.3% of the total, have more than 3 registered transponders.

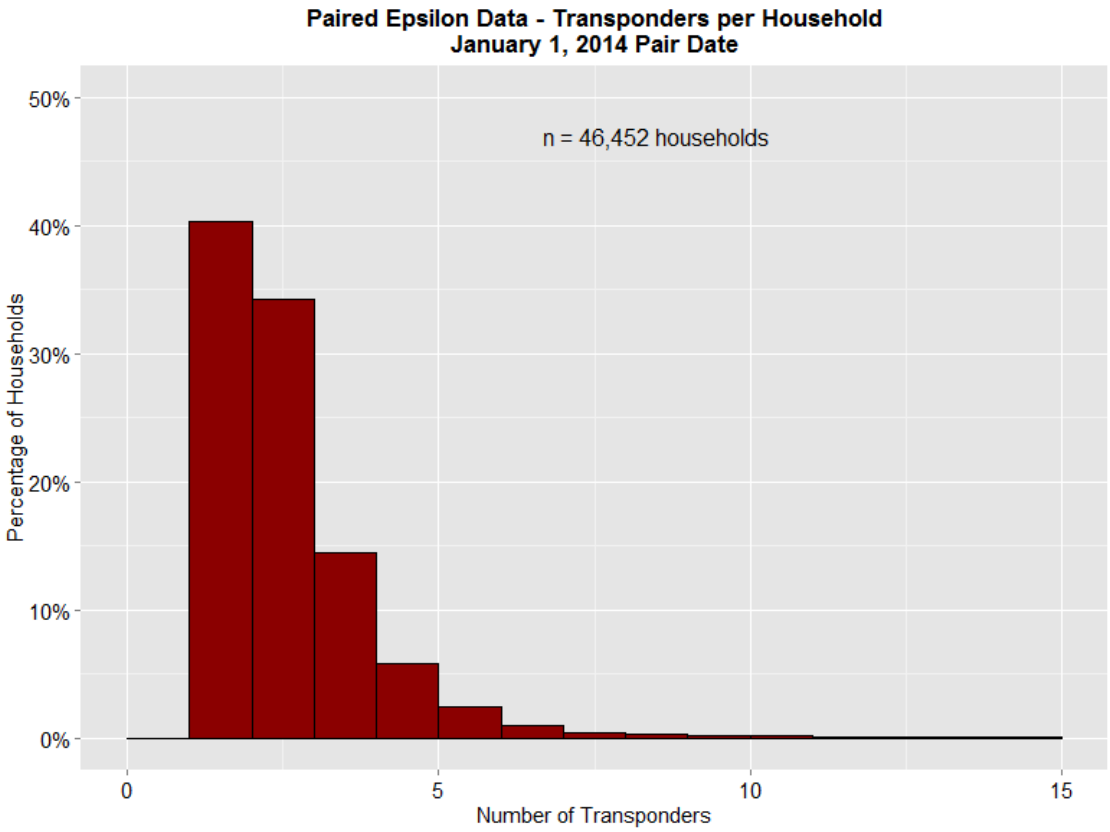


Figure 61: Distribution of Plates per Household in SRTA-Epsilon Matching Dataset

Figure 62 through Figure 67 illustrate distributions of selected Epsilon demographic variables in the SRTA-Epsilon matched dataset. Figure 62 shows the household income distribution in the matched sample. The average household income is significantly higher than the median income (more than $20,000 higher). The Epsilon mean is $2,817 higher (3.42%) than the $82,381 mean household income figure for the City of Atlanta, as reported by the 2009-2013 5-year American Community Survey estimates. The median household income also exceeds that of the ACS estimates for Atlanta by $15,869 (U.S Census Bureau, 2013). Also notable is the presence of lower-income households in the paired data set. While some of the income categories include very few households (in particular, the category from $15,000-$19,999), all of them include at least some matches. The two income categories ranging from $50,000 to $99,999 annually include a plurality of households, representing over 40% of the total.

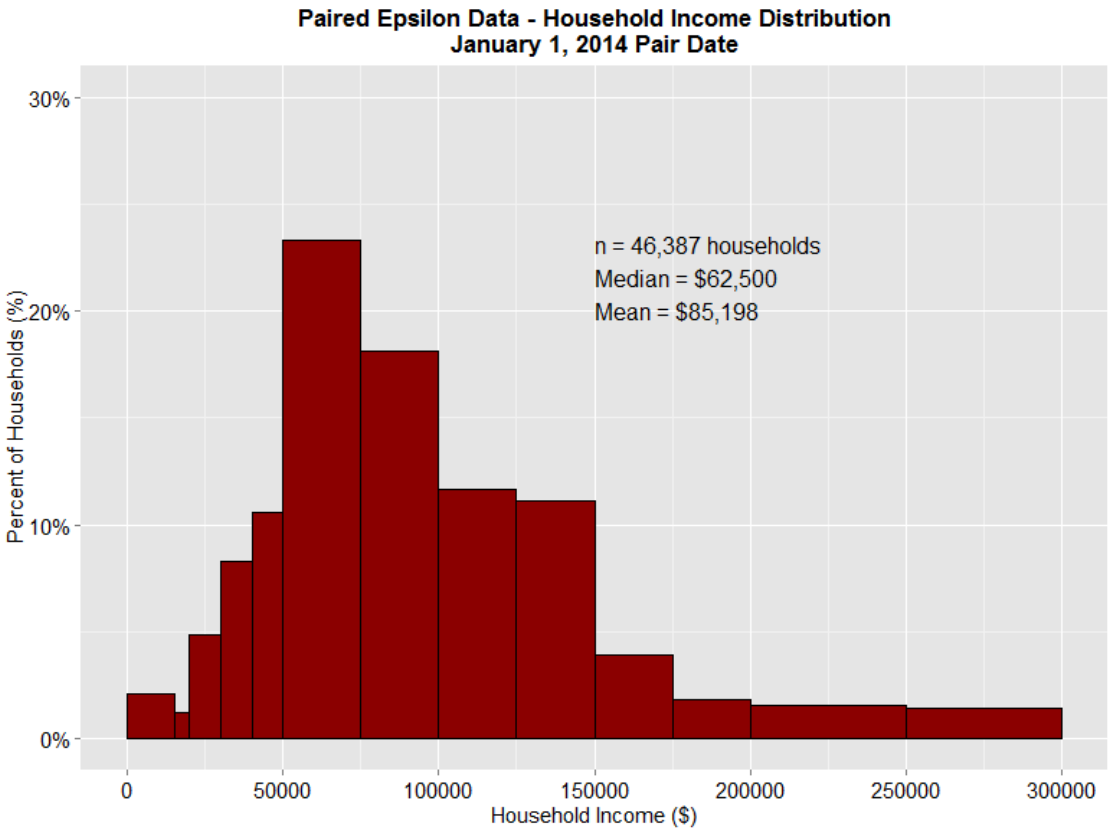


Figure 62: Household Income in SRTA-Epsilon Matching Dataset

Figure 63 illustrates the household size distribution in the matched data set; here, just over 30% of households have one member while households with two and three members make up a combined 40% of the sample. The ACS data for Atlanta report 45.9% of households as having a single member, while 29.3% of households are 2-person households and 11.5% are 3-person households (U.S Census Bureau, 2013). Again, this comparison involves different geographies as well as different data sources; a more direct comparison involving census results in the I-85 commutershed can be found in the Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter.

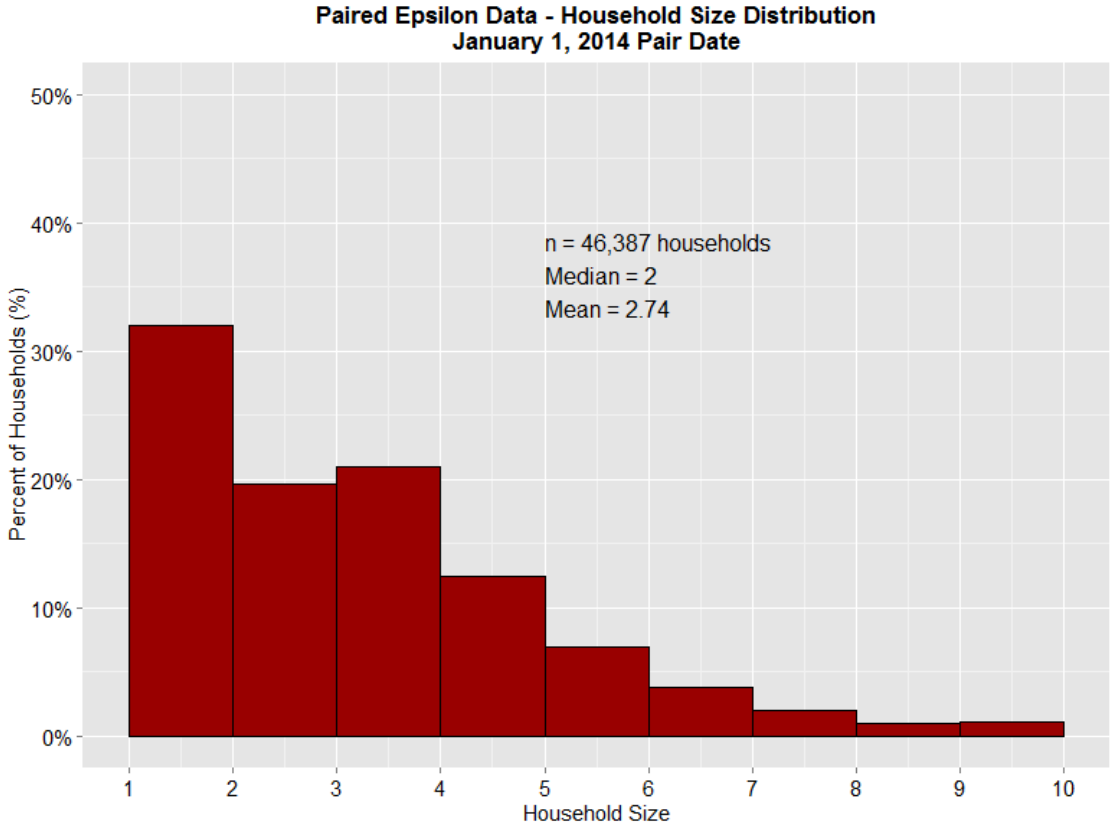


Figure 63: Distribution of Household Sizes in SRTA-Epsilon Matching Dataset

Figure 64 shows that over 80% of households in the matched sample are have at least some college education, while the proportion of households that have some high school is marginally smaller than the proportion that has attended graduate school. Within the ACS estimates for individuals 25 and over, the proportion of high school graduates is similar at 20.3%. The remaining categories differ more substantially. An estimated 31.6% of the ACS sample has an Associate's or Bachelor's degree, while over 50% of the matched Epsilon sample households have college degrees. The number of graduate degree holders reported by the ACS is much higher: 19.4% of the over-25 City of Atlanta population is estimated to have a graduate or professional degree (U.S Census Bureau, 2013).

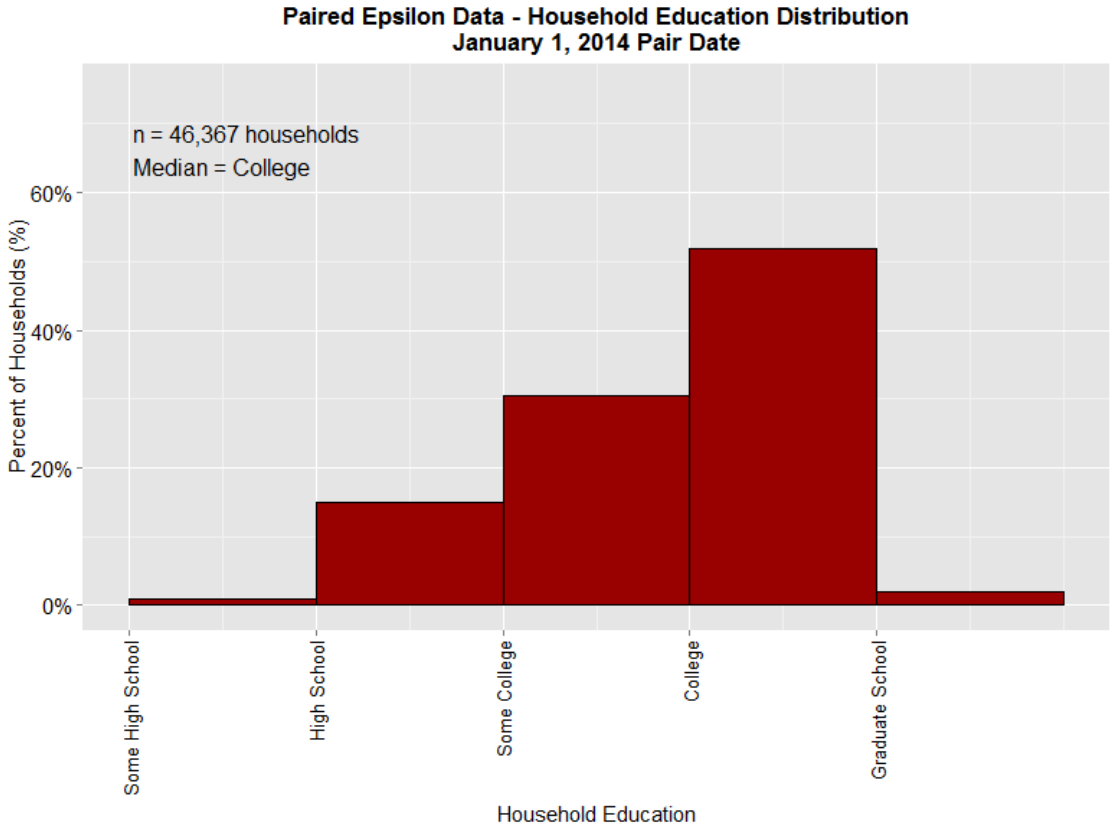


Figure 64: Household Education Levels in SRTA-Epsilon Matching Dataset

Figure 65 presents the head of household age distribution for the matched Epsilon sample. A plurality of households, over 30%, have a head of household between the ages of 35 and 44 years old. Very few fall within the 18-24 category, and roughly 10% are over the age of 65. That figure very closely resembles the 10% of the City population that is 65 or older as estimated by the ACS. Other categories do not align as neatly: the ACS estimates that 14.9% of individuals in Atlanta fall between 35 and 44 years of age. The ACS also estimates 19.6% of its sample is between 25 and 34 years old, while the matching Epsilon data report fewer than 15% of households of this age range. This raises another manner in which the comparison is not direct, however, as the Epsilon data reports household numbers while the ACS estimates are reported at the individual level (U.S Census Bureau, 2013).

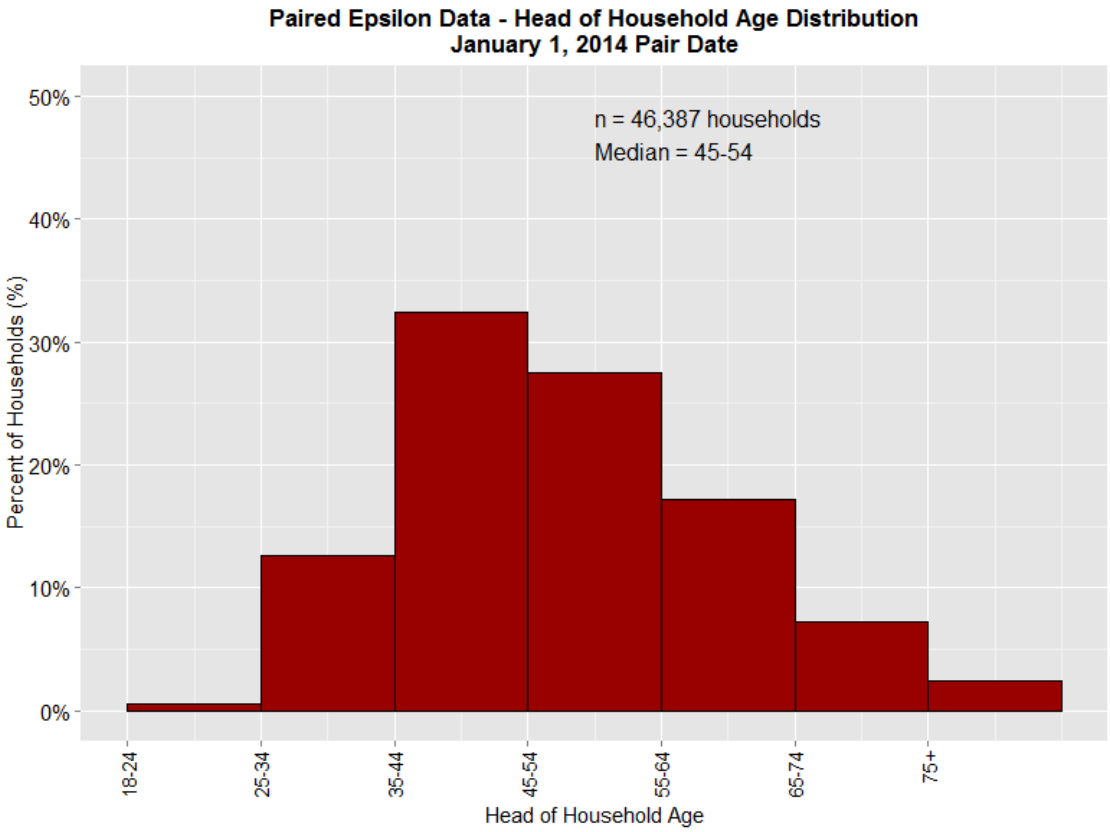


Figure 65: Head of Household Age in SRTA-Epsilon Matching Dataset

Figure 66 shows the number of households in the matching Epsilon data that fall into each of the marketing firm's home ownership categories. Of particular interest is the very small number of households that are renters; this is even more striking when one considers that the vast majority of households in the marketing data renter categories are only defined as "Probably renters." Within the ACS City of Atlanta estimates, over half of the households (54.6%) are identified as renter-occupied (U.S Census Bureau, 2013).

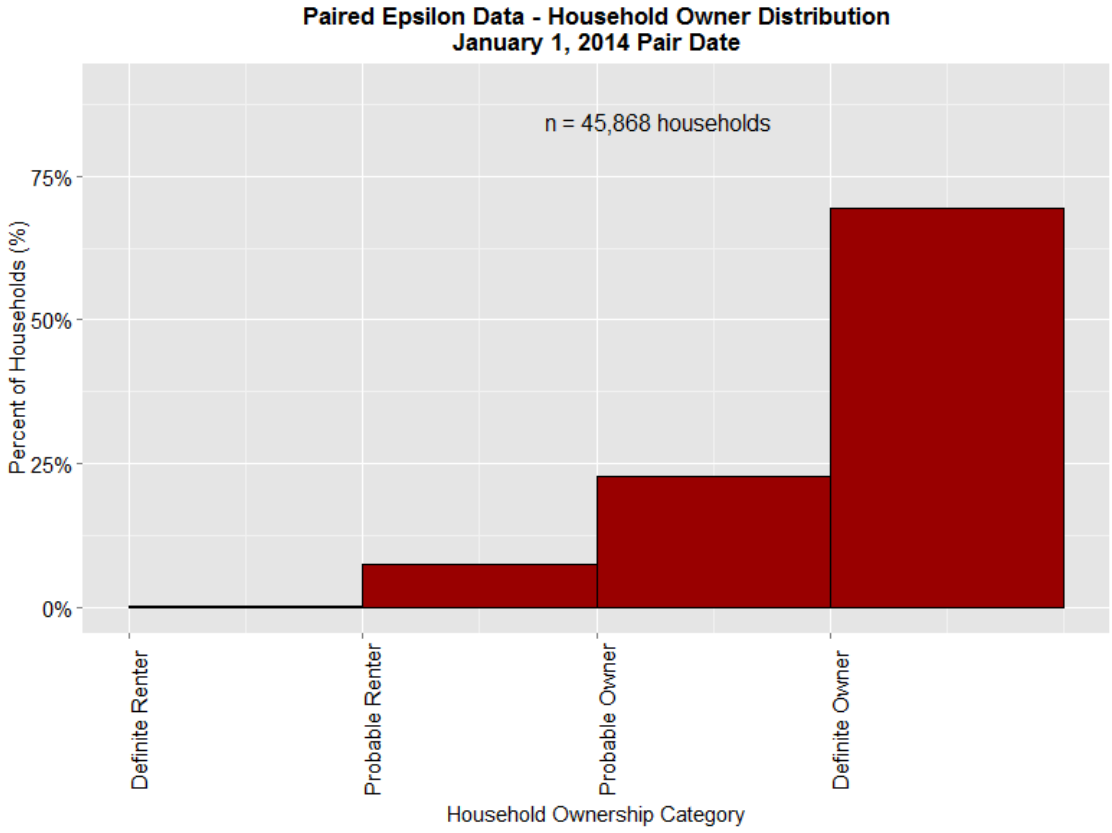


Figure 66: Home Ownership in SRTA-Epsilon Matched Data

The final chart, Figure 67, shows the potential dwelling types in the marketing data and the proportion of matched households that fall into each category. The great majority of the matched Epsilon households live in single-family dwelling units. The American Community Survey estimates for City of Atlanta households identify 40.2% of households as "1-unit, detached" and 5.3% as "1-unit, attached." 28.6% of the households in the ACS data include 20 or more units (U.S Census Bureau, 2013). Again, this discrepancy is likely the result of geographical differences rather than sample bias, as the ACS data include the whole City of Atlanta while the Epsilon marketing purchase was concentrated on the I-85 commutershed outside the I-285 perimeter of the City.

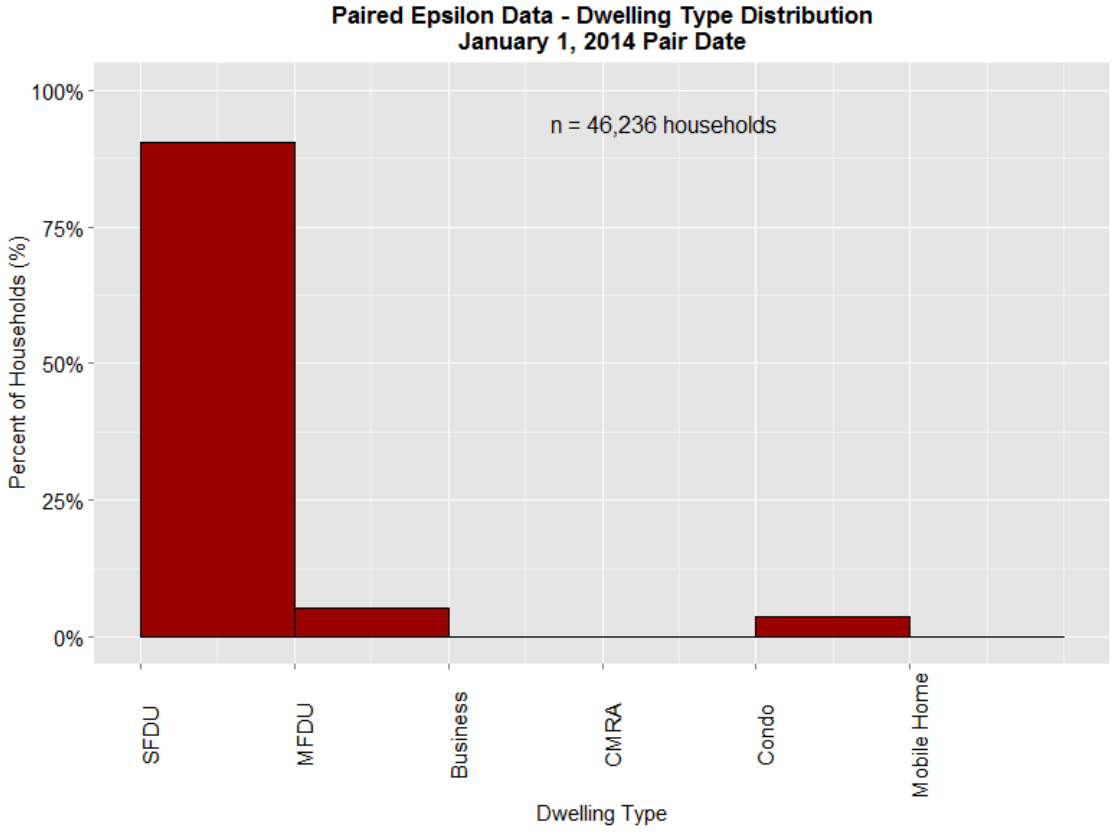


Figure 67: Dwelling Type in SRTA-Epsilon Paired Data

The distributions point to a number of differences between the Epsilon and Census data, or more accurately, between the I-85 commutershed households and the City of Atlanta population. Mean household incomes reported by the matched Epsilon set and the ACS are similar, though the median income value is much higher in the Epsilon data. The ACS also reports higher estimates of single-occupant households. The ACS education estimates include fewer individuals with undergraduate college degrees, but many more with graduate degrees. Far more home owners are present in the Epsilon data versus the ACS estimates. Similarly, the large proportion of renters (over 50%) in the ACS estimates are virtually unseen in the matched Epsilon data. It bears repeating that these comparisons are not perfect, as the scope of the ACS data differs geographically from that of the Epsilon data purchase; furthermore, the ACS data are reported at the individual level, rather than the household level. A more direct comparison of the Epsilon and ACS data, restricted to similar geographies, can be found in Chapter 7, Potential Sample Bias in Paired Vehicle Activity and Marketing Data.

Comparison between Paired and Overall Households

The Epsilon households that were successfully paired with SRTA Peach Pass transponders represent a fraction of the total households in the demographic data set: the 46,452 households represent 13.3% of the purchased Epsilon sample. The third chapter of this dissertation, entitled Data Sources, provides an overview of the full Epsilon population. Table 20 presents the results of a comparison between that full population and the narrower paired sample that was discussed in the previous sections of this chapter.

Table 20: Comparison of Paired Households and All Epsilon Households

	Full Epsilon Data Set	Paired Epsilon Sample
Number of Households	349,134	46,452
Household Income		
Mean	$61,862	$85,198
Median	$62,500	$62,500
25th Percentile	$35,000	$45,000
75th Percentile	$87,500	$112,500
Skewness	1.56	1.21
Kurtosis	6.93	5.01
Mann-Whitney Test Result	p<2.2x10-16	
Household Size		
Mean	2.40	2.74
Median	2	2
25th Percentile	1	1
75th Percentile	3	4
Skewness	1.47	1.14
Kurtosis	5.39	4.16
Mann-Whitney Test Result	p<2.2x10-16	
Household Education		
Mean	3.04	3.39
Median	3	4
25th Percentile	2	3
75th Percentile	4	4
Skewness	-0.22	-0.69
Kurtosis	2.06	2.69
Mann-Whitney Test Result	p<2.2x10-16	
Head of Household Age		
Mean	3.72	3.80
Median	3	4
25th Percentile	3	3
75th Percentile	5	5
Skewness	0.56	0.46
Kurtosis	2.83	2.79
Mann-Whitney Test Result	p<2.2x10-16	

Table 20 provides summary statistics for a subset of the demographic variables in the Epsilon data set. The table highlights a number of notable differences between the paired data subset and the full data purchase. Household income levels are higher on average in the paired subsample, though the median values are the same. This is further reflected in the 25th and 75th percentile values, both of which are higher in the paired sample. The skewness values indicate that both distributions are right-trailed, while the

kurtosis results indicate that both are more peaked than the normal distribution. The full sample is less symmetric and more peaked than the paired sample.

The household size variable follows a similar pattern: the paired sample includes a higher proportion of larger households than the full data set. This is reflected in the higher mean and 75th percentile values. Like the household income distributions, the household size distribution is less symmetric and more right-tailed in the full data set. The full data set is also more peaked than the paired sub-sample. The household education variable distributions differ slightly in that both are left-tailed; in this case, the paired data are less symmetric. The higher mean and 75th percentile education values in the paired data support this as well.

Out of the four variables compared here, the head of household age measure is the most similar across the two data sets. Though the median value is a full unit higher in the paired sample (median of 35-44 in the full sample, median of 45-54 in the paired sample), the mean values differ only slightly, and the skewness and kurtosis measures are similar as well. For each variable under examination, researchers used the Mann-Whitney two-tailed test to compare the distributions of the full sample and the paired sample. In each case, the resulting p-value was virtually indistinguishable from zero; the null hypothesis of distributional equality was rejected.

Chapter Overview

The process of pairing the SRTA lane use data with the Epsilon demographic data involves many steps that each include the potential for data loss and bias. Foremost among these is the final stage in the process, in which the address-matched SRTA records attempt to find matches in the Epsilon data set. This step narrows the transponder sample to roughly one-third of the GTRI-matched records and one-fifth of the total records. Prior to that, however, complications in the structure of the SRTA data restrict the scope of users that survive the pairing process. Restricting the households to those within the I-85 commutershed improves the match rate between GTRI-matched households and Epsilon households, but removes over 6,000 transponders from the final sample. The resulting sub-sample of demographic data exhibits notable differences from the complete sample: larger households, more higher-income households, and more highly educated households. A brief comparison with the Census Bureau's estimates of related measures from the City of Atlanta points to significant differences with that data set as well, especially in the areas of home ownership and dwelling type (single family versus multi-family). This issue is further explored in Chapter 7, Potential Sample Bias in Paired Vehicle Activity and Marketing Data, which investigates household demographic comparisons at different stages in the pairing process using commutershed-restricted Census data.

CHAPTER 6

DATA QUALITY AND TREATMENT

During the course of working with the SRTA Express Lane use data and the Epsilon marketing demographic data, a number of issues arose with the structure and quality of the data that may have affected the match rate between the two data sources and initial analytical results that arose out of the matched dataset. This chapter will describe those issues and the methods used to address them. The first section describes the problem with the structure of the SRTA Account data, and the second does the same for the SRTA Trip stream data. The next section discusses the time series relationships of plates and transponders in that Trip data stream. This is followed by an investigation into the stability of the individual and combined data sets over the course of the three years of analysis. The chapter then examines issues with the Epsilon marketing data and the revised data set that attempted to correct those issues. Finally, the chapter ends with a look at the quality of the SRTA transponder detection data that serves as the foundation for much of the analysis in this dissertation.

Account Transponder and Vehicle Issue

As part of the process of pairing Express Lane use data to household demographic data, researchers needed a link between the Peach Pass transponder identifier and the license plate of the vehicle in which that transponder was used. Using the license plate data, researchers at GTRI pulled addresses from the Georgia vehicle registration database via a single-blind process. Addresses records were returned to Georgia Tech without the

license plate numbers and names for privacy purposes. The author then attempted to pair the addresses with the Peach Pass transponders, so that the demographic data could be paired with the SRTA lane use data.

Because the transponder-to-address pairing process required a license plate, researchers could only find addresses for transponders that had associated license plates. Transponders without license plate data needed to be excluded from the demographic analyses for this reason. Similarly, transponders with too many license plates were also initially excluded from demographic data pairing. If a transponder were to be associated with multiple license plates, those plates may match to multiple household addresses, potentially making it impossible to identify which set of household data applies to a transponder.

Unfortunately, the structure of the SRTA-provided Peach Pass Account data yielded many instances of this situation. The Account data lacked a join table linking every transponder to one-and-only-one vehicle/license plate. This was not an issue when an account had only one transponder and one license plate; in that case the pairing was obvious. Many accounts, however, had multiple transponders and/or license plates. Because of the lack of a linking element, researchers could not identify in these cases which transponder was paired with which license plate. In situations where the license plates in an account were matched with different household addresses, there was no way to identify which address the Peach Pass transponders in that account were paired with.

To examine the scope of this and other structural problems, Table 21 presents a summary of the SRTA Account stream data from January 1, 2014. The table shows the various issues present in the structure of the Account table. Registered transponders and

vehicles have, in different instances, one-to-one relationships, one-to-many relationships, and many-to-one relationships. Nearly 50% of the active accounts have active transponders with a one-to-one relationship with a vehicle; that is, there is only one transponder and one vehicle registered to that account. The next largest group of accounts, over 38% of the total, has multiple vehicles (greater than or equal to two) and transponders. Notably, the transponders associated with these accounts far outnumber those with a one-to-one relationship with a vehicle. The 286,066 transponders in the many-to-many set are more than double the 131,494 in the one-to-one set. These transponders cannot be paired with a specific vehicle due to the lack of a joining element. A trivial number of accounts have registered transponders without any vehicles; a similarly low proportion of accounts have just one registered vehicle and more than one active transponder. The other possible many-to-one relationship, in which an account has just one registered transponder but multiple vehicles, occurs in 6.3% of the active accounts. The table entries in parentheses refer not to accounts but to transponders. The two bolded entries at the end of the table indicate the scope of the scope of the problem of transponders associated with multiple accounts; a total of 610 transponders on January 1, 2014 fell into this category.

Table 21: Account Data Breakdown

Account Stream Data – 01/01/2014	**Accounts**	**% of Total**
Active Accounts in SRTA Data (Status = A, I, P)	278,170	100%
Active accounts with one active transponder and one vehicle	131,494	47.27%
Active accounts with one active transponder and no vehicles	3	0.0011%
Number of transponders within these accounts	(3)	N/A
Active accounts with one active transponder and multiple vehicles	17,531	6.30%
Number of transponders within these accounts	(17,531)	N/A
Active accounts with no active transponders and no vehicles	18,808	6.76%
Active accounts with no active transponders and one or more vehicles	3,528	1.27%
Active accounts with multiple active transponders and one vehicle	165	0.059%
Number of transponders within these accounts	(331)	N/A
Active accounts with multiple active transponders and multiple vehicles	106,641	38.34%
Number of transponders within these accounts	(286,066)	N/A
Number of transponders associated with two active accounts	**(608)**	N/A
Number of transponders associated with three active accounts	**(2)**	N/A

Express Lane Trip Stream Issues

The Express Lane Trip data stream, which provides a daily summary of toll lane trips, was originally thought to be a partial solution to the one-to-many and many-to-many relationships between account transponders and vehicles discussed in the previous section. The vast majority of trip records, over 99.9%, provide both a transponder identifier and a vehicle plate number, theoretically allowing for the one-to-one join

between transponders and plates that is missing from accounts with multiple instances of either element. This assumption was used in some preliminary analyses. Further investigation of the Trip stream data, however, revealed that the relationship was more complicated and that the transponder-vehicle pairs could not be used without further scrutiny. The first issue was the lack of a true one-to-one pairing between the transponder and vehicle elements in the Trip stream. An examination of the trip data from November, 2011 through December, 2014 revealed many instances in which transponders were associated with multiple license plates and license plates were associated with multiple transponders. Similarly, the data had records where no transponder was reported, no plate was reported, or both fields were empty. Table 22 presents an overview of the transponder side of these Trip stream transponder-plate relationships.

Table 22: Transponder-Plate Relationships in Trip Stream Data

Trip Stream Transponders: 11/2011 – 12/2014	Transponders	% of Total
Total Unique Transponders in Trip Stream	254,251	100%
Transponders with 1:1 Plate Matches	194,326	76.4%
Transponders associated with multiple plates, plate associated with one transponder	14,701	5.8%
Transponders associated with one plate, plate associated with multiple transponders	28,774	11.3%
Transponders associated with multiple plates, plates associated with multiple transponders	7,548	3.0%
Transponders with no plates associated	8,902	3.5%

Figure 68 provides a breakdown of the different categories of transponder and plate pairings in the SRTA Trip data. The data listed over 850,000 unique license plates,

though over 400,000 of those were never paired with transponders. A much smaller number of transponders, 8,902, were never paired with license plates. Nearly 200,000 unique transponder and plate pairs appear in the Trip data. These were transponders that were only ever associated with a single plate, and plates that were only ever associated with a single transponder. The remaining counts address transponders that were paired with multiple plates, plates that were paired with multiple transponders, and many-to-many relationships in which both of these situations occurred.

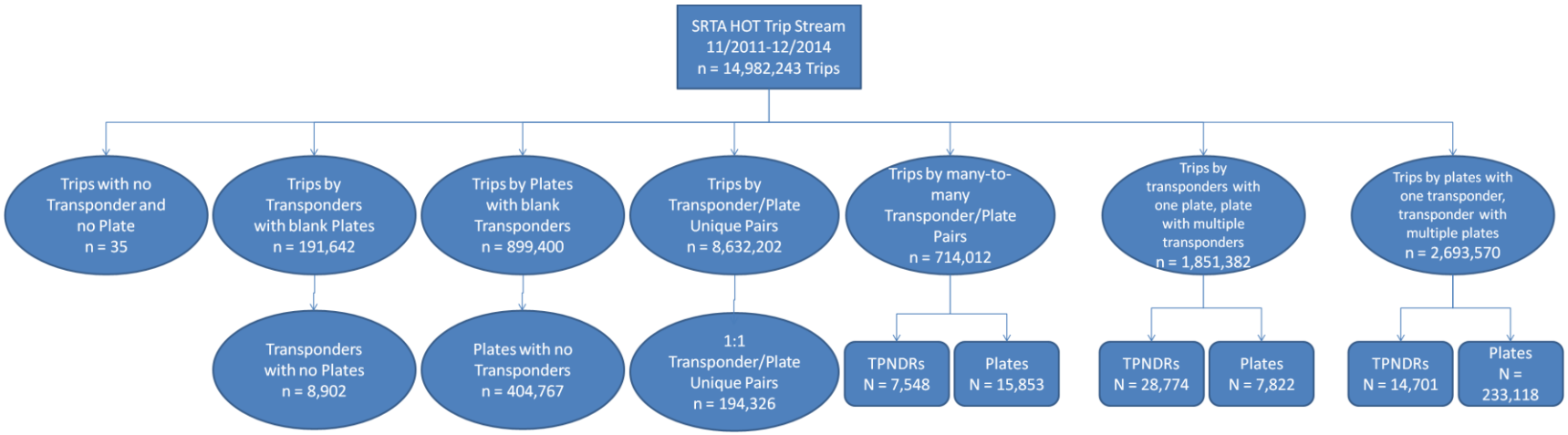


Figure 68: Trip Stream Plate and Transponder Breakdown

A very small percentage (0.0002%) of Express Lane trip records had no transponders or license plate data. Trips with transponder identifiers but no plate numbers make up 1.28% of the trip set. Trips by transponders associated with multiple plates account for 17.98% of the total; these transponders are discussed more in the next section. Nearly 900,000 trips, making up 6.00% of the total, include license plate data but no transponder identifier. These are likely violation trips by vehicles without Peach Passes. The majority of trips, 57.62%, includes both transponder and plate data, and represent unique pairings of those two elements. The Trip stream also includes transponders that are paired with multiple plates, while those plates are also associated with multiple transponders. This many-to-many relationship between transponders and plates accounts for 4.77% of the total trip count. Finally, trips by transponders associated with a single plate, while that plate is associated with multiple transponders over the 38 month timeframe, make up 12.36% of the total.

Table 23 shows the vehicle plate aspect of the transponder-plate relationship in the SRTA Trip stream data. Almost 23% of the total plates counted over 38 months have a unique transponder paired with them, while nearly 50% have no associated transponder. While less than 2% of all plates have a many-to-many relationship with Peach Pass transponders, over 25% of the plates have a one-to-many relationship, in which their associated transponders are also tied to other license plates.

Table 23: Plate-Transponder Relationships in Trip Stream Data

Trip Stream Plates: 11/2011 – 12/2014	Transponders	% of Total
Total Unique Plates in Trip Stream	855,886	100%
Plates with 1:1 Transponder Matches	194,326	22.71%
Plates associated with multiple transponders, transponders associated with one plate	7,822	0.91%
Plates associated with one transponder, transponder associated with multiple plates	233,118	27.24%
Plates associated with multiple transponders, transponders associated with multiple plates	15,853	1.85%
Plates with no transponders associated	404,767	47.29%

The overall effect of this database structure issue is to narrow the potential pool of Peach Pass users that can be studied using the available demographic data. This effect can also be seen in the overall match rates between the SRTA lane use data and the Epsilon marketing data. Later, this dissertation will discuss how the narrow sample of one-to-one transponders was expanded to include some fraction of the remaining Peach Passes and vehicles.

Time Series of Transponder and Plate Relationships

One element of the transponder-to-license plate relationships in the Trip stream that appeared after further investigation was the overlapping nature of the pairings. After identifying the first and last detection of a unique transponder and plate pair, researchers discovered that the same transponder was often associated with multiple plates within the same time interval. That is, rather than one transponder cleanly transitioning from one vehicle to another, instead it would appear to be associated with two different vehicles concurrently. This may have been an artifact of the method by which the operating firm ETTC reported vehicle plate numbers alongside the transponder identifiers. Researchers

suspected that the firm used faulty database joins that may have yielded incorrect license plates for accounts with multiple vehicles, or image recognition software that may have reported variations on the same license plate in different instances.

A script that examined over three years' worth of Express Lane trips, from November 2011 through December 2014, recorded the first and last instance of each unique transponder and plate pairing. The script then reported the total number of unique transponders detected within that timeframe, and also the number of transponders associated with multiple license plates within the same timeframe. That is, transponders whose plate pairings overlapped. The script found 882,850 total unique transponders over the 38 month timeframe. Of those, 4,772 transponders were detected within overlapping plate-pairing intervals. For these 4,772 transponders, which make up 0.54% of the unique transponder population, the SRTA Trip summary data cannot reliably be used to pair them with a unique license plate.

A second script read the same Express Lane trips and identified transponders that were associated with more than one plate in the Trip summary data. This script looked for Peach Passes that were detected at least 250 times within the 38 month timeframe from November 2011 through December 2014. It generated a list of toll lane trips, ordered sequentially by date and time, along with the transponder identifier and the license plate associated with that specific trip. From this, researchers could see when individual transponders changed the plate with which they are linked. A variation of this script also included blank values for the license plate field, thus identifying transponders that were linked to at least one license plate and to blank license plate records. Figure 69 below illustrates one example from September 2014 of such a transponder. The

transponder is first linked to ‘CCXXXXX,’ then ‘BSXXXXX.’ The link switches between one plate and the other multiple times throughout the month. The first occurrence occurs during the workday on October 13th; the 5:38AM trip occurs with one license plate, while the 6:30PM trip occurs with the other. Running this script for one month, October 2014, identified 328 transponders that fit the given criteria when blank license plates were included. Without blank license plates, 186 transponders matched the criteria (at least two plates and at least 20 trips). Note that the script in this instance was looking for 20 trips within the month of October; running the script for all 38 months would expand the interval for the 20 trip criteria and would thus include more transponders. The 20 trip criteria was increased to 250 for the 38 month duration. The minimum trip criteria was computationally necessary to allow the script to run. Note that the license plates in the image below have been masked for privacy purposes.

transponderID	plateNumber	detectionTime
329959	CC[redacted]	10/1/2014 18:36
329959	CC[redacted]	10/2/2014 7:11
329959	CC[redacted]	10/2/2014 18:22
329959	CC[redacted]	10/6/2014 7:01
329959	CC[redacted]	10/6/2014 17:20
329959	CC[redacted]	10/7/2014 7:08
329959	CC[redacted]	10/8/2014 18:35
329959	CC[redacted]	10/13/2014 5:38
329959	BS[redacted]	10/13/2014 18:30
329959	CC[redacted]	10/14/2014 5:36
329959	CC[redacted]	10/14/2014 15:10
329959	BS[redacted]	10/15/2014 9:34
329959	BS[redacted]	10/15/2014 14:22
329959	CC[redacted]	10/20/2014 6:02
329959	CC[redacted]	10/20/2014 15:18
329959	CC[redacted]	10/21/2014 5:58
329959	CC[redacted]	10/22/2014 6:01
329959	CC[redacted]	10/22/2014 16:40
329959	CC[redacted]	10/24/2014 6:02
329959	CC[redacted]	10/24/2014 15:40

Figure 69: Example Transponder Associated with Two Plates

Running the script for the full 38 months yielded 298 frequently used (at least 250 trips) transponders that were associated with multiple plates, not including blanks among the possible license plate values. A total of 579 unique license plates were associated with those transponders. The vast majority of these transponders, 96.3%, were associated with two different license plates. The remaining transponders were associated with three different license plates.

This transponder-to-license-plate pairing issue has implications both for the design of the database in which the lane use data are stored and for this dissertation. It may be the case that a user is switching the transponder from one vehicle to another, in which case the data are correct. If not, it may be the case that the database is querying the license plate records incorrectly and returning faulty license plate data. The presence

and use of license plate cameras on the corridor also complicates the data, as it is evident that some of the transponders are associated with two 'different' license plates that differ by very few characters. Figure 70 provides an example of this. This indicates that some faulty image recognition may be the cause of the differing license plate results.

	A	B
1	transponderID	plateNumber
2	00187572	[redacted]ROB
3	00187572	[redacted]R0B
4	00037467	VOL[redacted]
5	00037467	V0L[redacted]

Figure 70: Similar License Plates in SRTA Data

For the purposes of the analyses presented here, the multiple plate pairing issue complicates the one-to-one join needed between transponders and license plates to properly tie demographic data to Express Lane use data. The result is that the Trip summary data does not provide a clean match, but rather resembles the Account data in that it includes many-to-many relationships between transponders and license plates. These issues further narrow the subset of Trip stream transponders and license plates that can be included in the analyses presented in this dissertation.

Sample Stability in SRTA Express Lane Data

One of the primary issues in the methods used in this dissertation involves the cross-sectional nature of some of the data sources and the longitudinal nature of the other data sources. The Epsilon demographic data and the registration database matching by GTRI are both cross-sectional in that they were the result of one-time queries that returned records from a database at only a single point in time. The Epsilon dataset was dated March 6, 2013, while the GTRI registration database matching was performed on May 23, 2014. Both of these data sets present their results from that day only.

In comparison, the SRTA Express Lane data tables are updated every day, and in some cases multiple times a day. The elements that are used to connect lane use data with registration and demographic data may potentially change every day. Vehicles may have been associated with an account prior to May 23, 2014, for example, and then removed from that account after that date. Similarly, a household that registered for a Peach Pass after March 6, 2013 would not appear in the SRTA-Epsilon paired dataset as that address would not have been included in the marketing data purchase. This section examines the stability of the SRTA lane use data, Epsilon demographic data, and GTRI registration data throughout the course of facility operations.

The total number of license plates in the STRA Express Lanes trip summary data from November, 2011 through December, 2014 is illustrated below in Figure 71. The figure also shows the number of new and dropped license plates each month. New plates are those which were not previously seen in the Trip summary data, and thus that month contains their first detection. Dropped plates are those whose last observation occurred in that month. Note that this chart is based solely on the Express Lanes Trip summary stream, so it only captures toll lane trips, not general purpose lane trips. The chart illustrates the 'churn' in the vehicle plate data: each month, thousands of license plates are detected for the first time. Each month also see thousands of license plates that are detected for the last time, at least through December 2014. An average of 75,578 total license plates are detected each month. Excluding the first month (in which all detected plates are 'new'), an average of 22,347 license plates are detected for the first time each month. An average of 20,613 plates are dropped each month (again, excluding the last month in which all remaining plates are 'dropped').

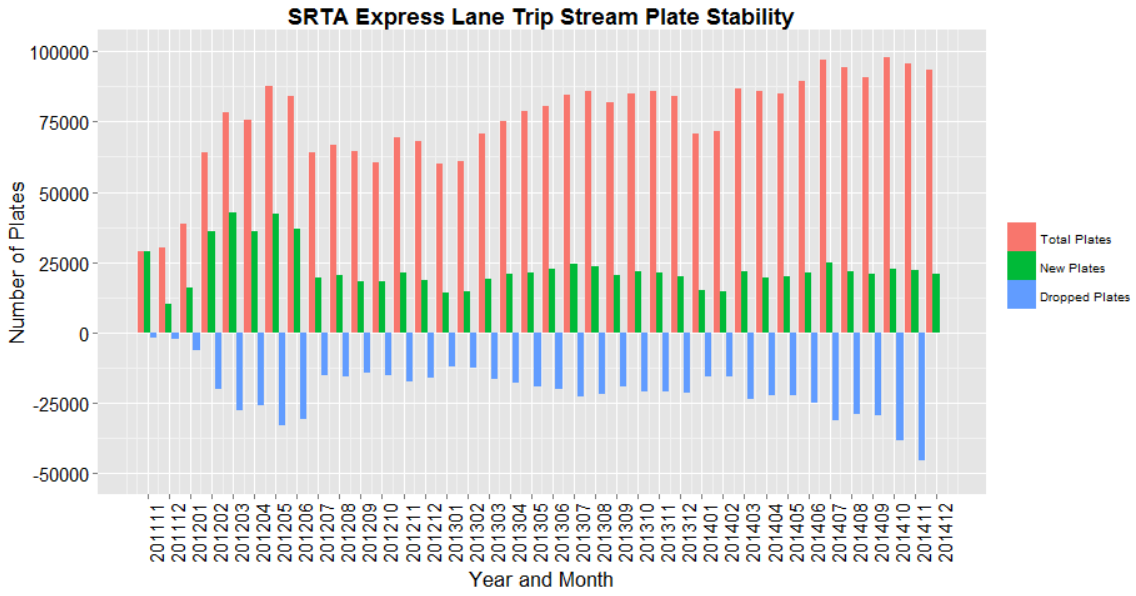


Figure 71: SRTA Trip Stream Plate Stability

Similarly, Figure 72 shows the new, dropped, and total Trip stream transponders each month. In this case, the average number of total transponders detected per month is 53,713. 6,082 new transponders are detected each month on average, while 4,962 are dropped.

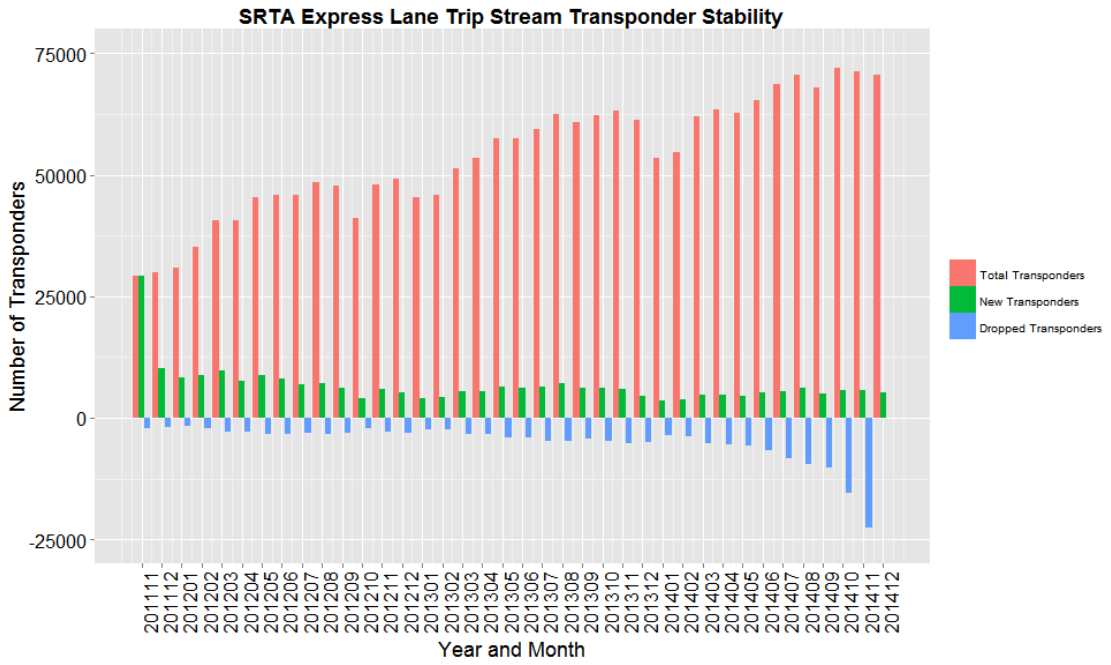


Figure 72: SRTA Trip Stream Transponder Stability

Figure 73 shows new and dropped transponders from the SRTA Vehicle detection data. This data source includes detections from both the Express Lanes and the general purpose lanes and so gives a more complete picture of the turnover in the data.

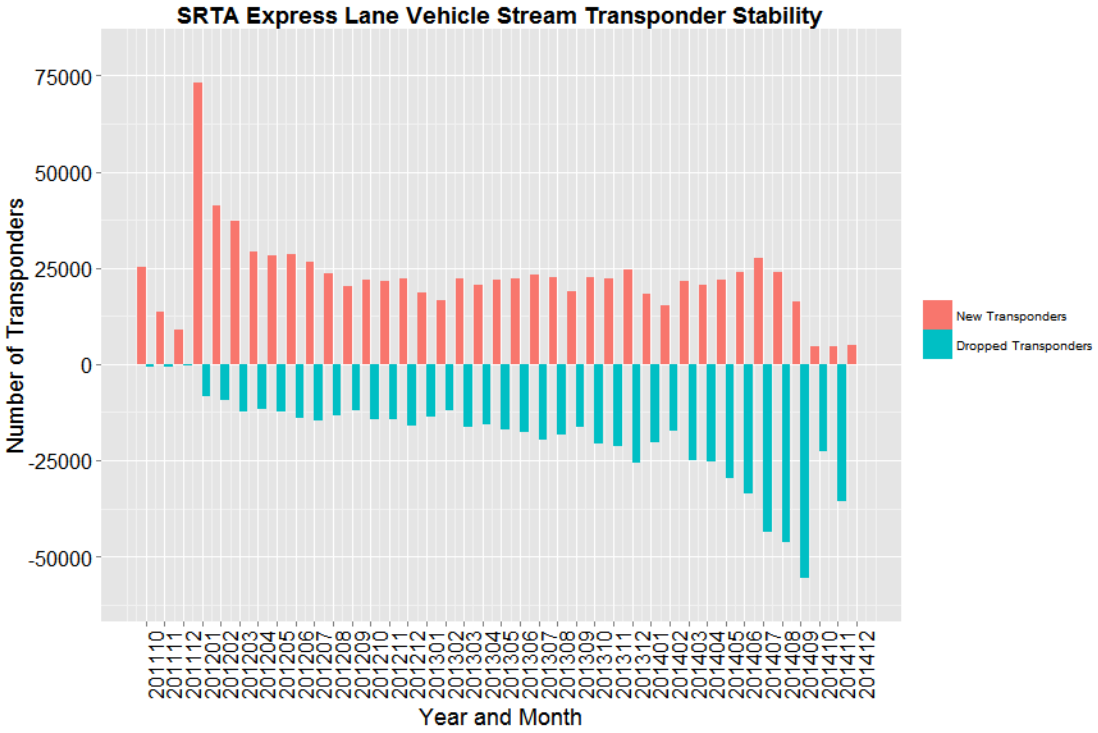


Figure 73: SRTA Vehicle Stream New and Dropped Transponders

Figure 74 illustrates the total counts of unique transponders identified in the Vehicle detection stream. The trend has been increasing steadily since early 2012; the inactive General Purpose lane detectors are likely the cause of the very low counts in the first three months. Had the time frame expanded into 2015, the drop at the end of the figure would be provided with more context to see if it represents an aberration or a change in the trend.

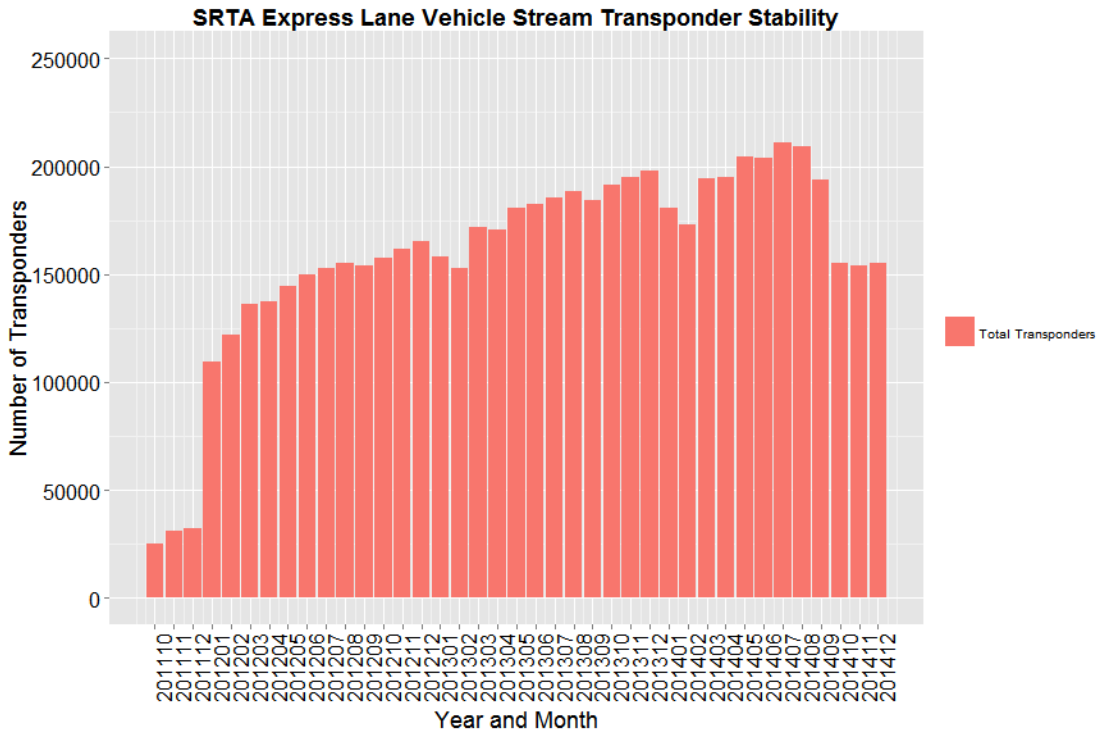


Figure 74: SRTA Vehicle Stream Total Transponders

The two figures above present only those plates and vehicles that were observed on the corridor; this may not provide a full account of the system's characteristics. The next set of figures uses the SRTA Account data to illustrate the total number of accounts as well as registered, rather than observed, vehicles and transponders over the thirty-eight month time frame. The figures also illustrate the new accounts as well as the new and dropped vehicles and transponders per month. Figure 75 illustrates the steadily-increasing numbers of total registered Peach Pass accounts from January, 2012 through December, 2014.

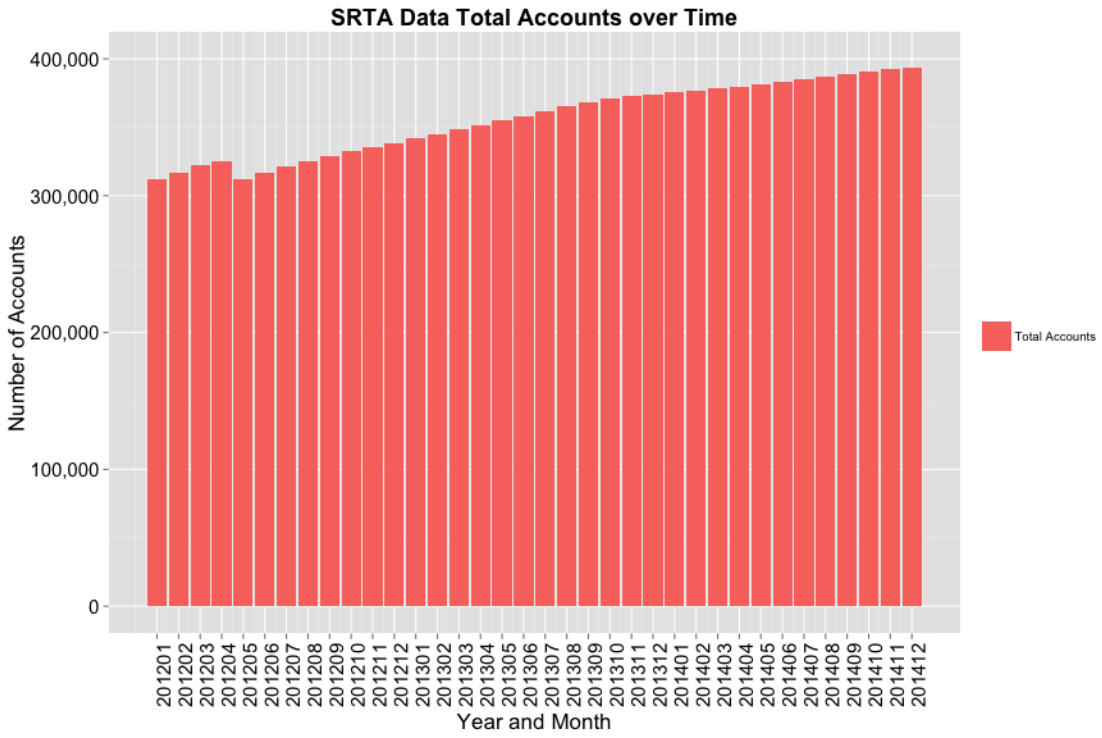


Figure 75: Total SRTA Accounts over Time

Figure 76 shows the total counts of SRTA accounts in Active status over those three years of operations. The trend is very similar to that of the previous chart; each month, the proportion of active accounts ranges from a minimum of 71.8% to a maximum of 75.2%.

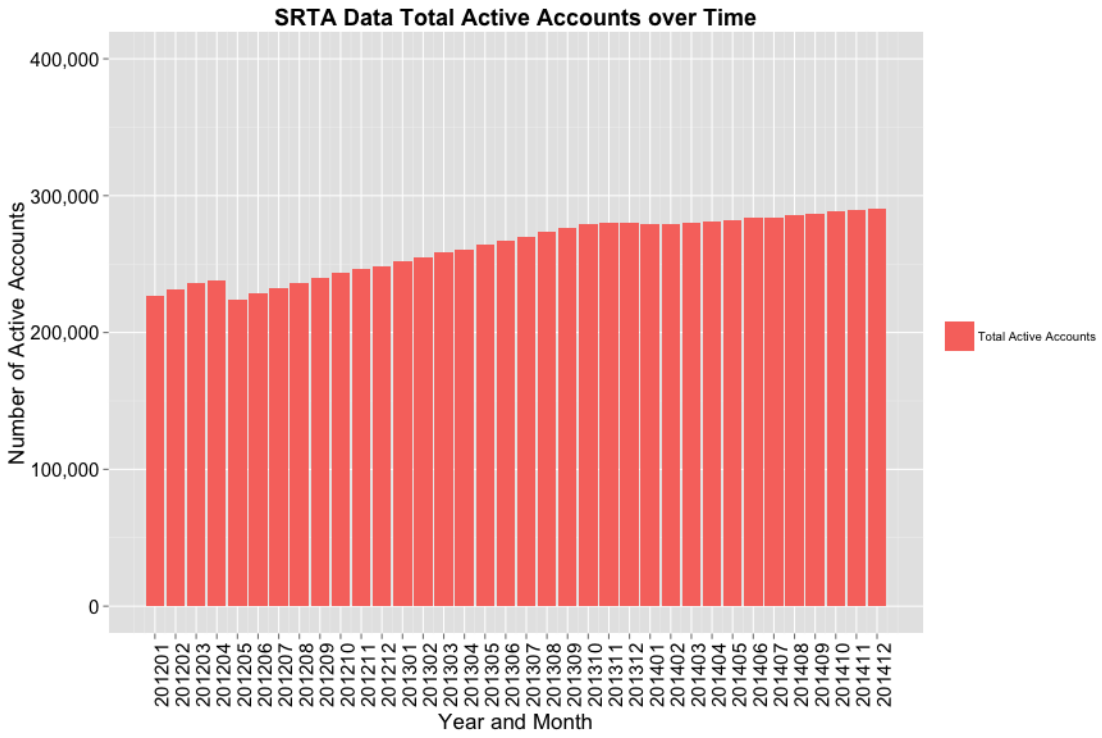


Figure 76: Total Active SRTA Peach Pass Accounts over Time

The rate of new Peach Pass account creation is shown in Figure 77. While the number of new accounts began to decrease towards the end of 2013, the system is still adding nearly 2,000 accounts per month. The drop in April of 2012 represents missing data in the Account stream; this issue is described in greater detail later in this chapter.

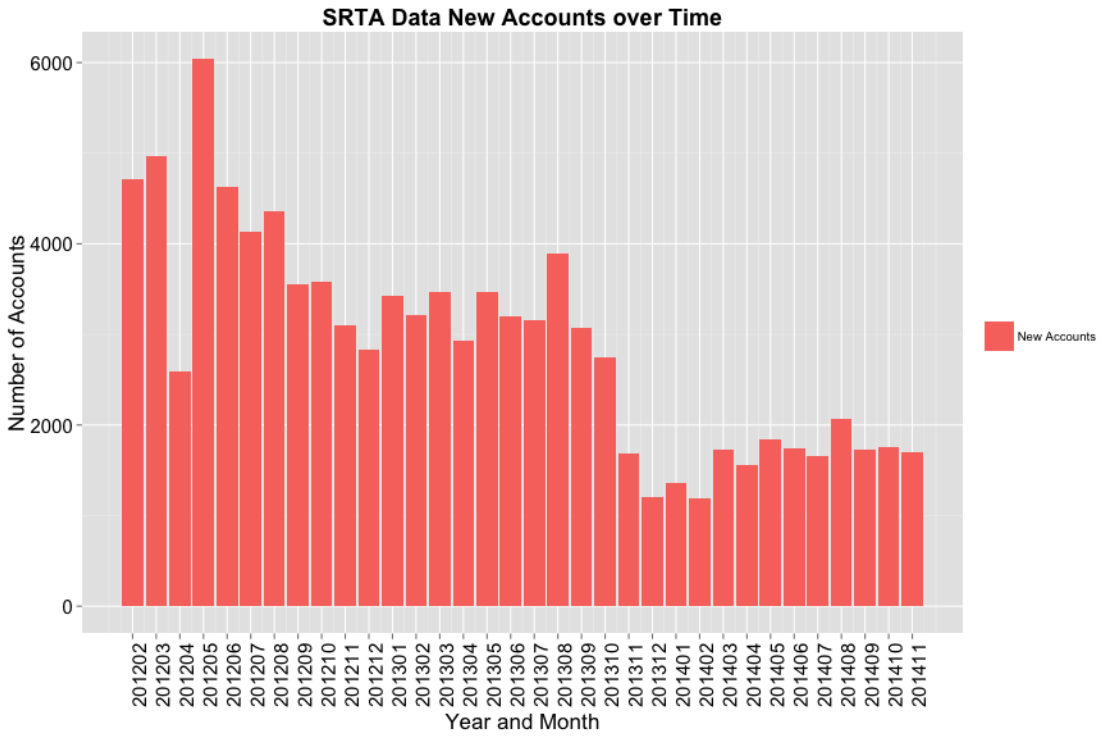


Figure 77: New SRTA Accounts over Time

Figure 78 focuses on the vehicles registered in the SRTA account data stream. For every month, the rate at which vehicles are added exceeds the rate at which accounts are created. Relatively few vehicles are removed from the account tables each month. This figure shows part of the difficulty in the demographic pairing method used in this study: as mentioned above, the query against the registration database was executed in May of 2014. Roughly 20,000 vehicles were added to the database after that month; none of those are included in the analysis.

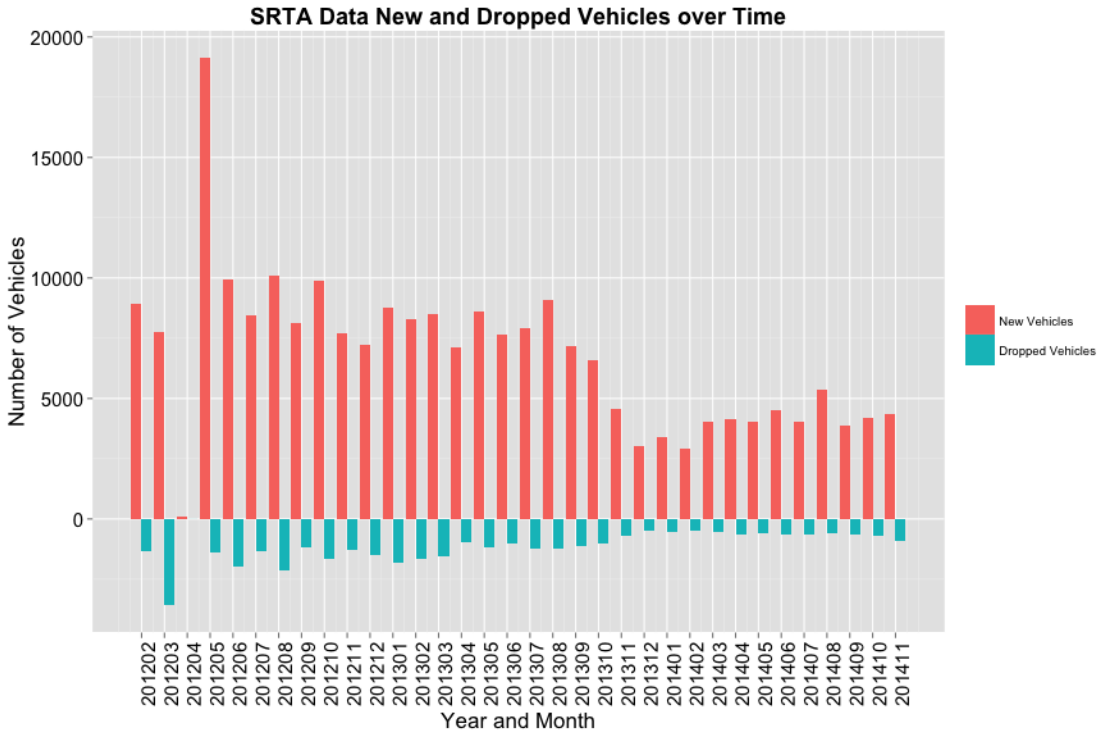


Figure 78: New and Dropped SRTA-Registered Vehicles over Time

Figure 79 shows the total count of registered transponders which, similar to the number of registered accounts, increases steadily over the three years under study.

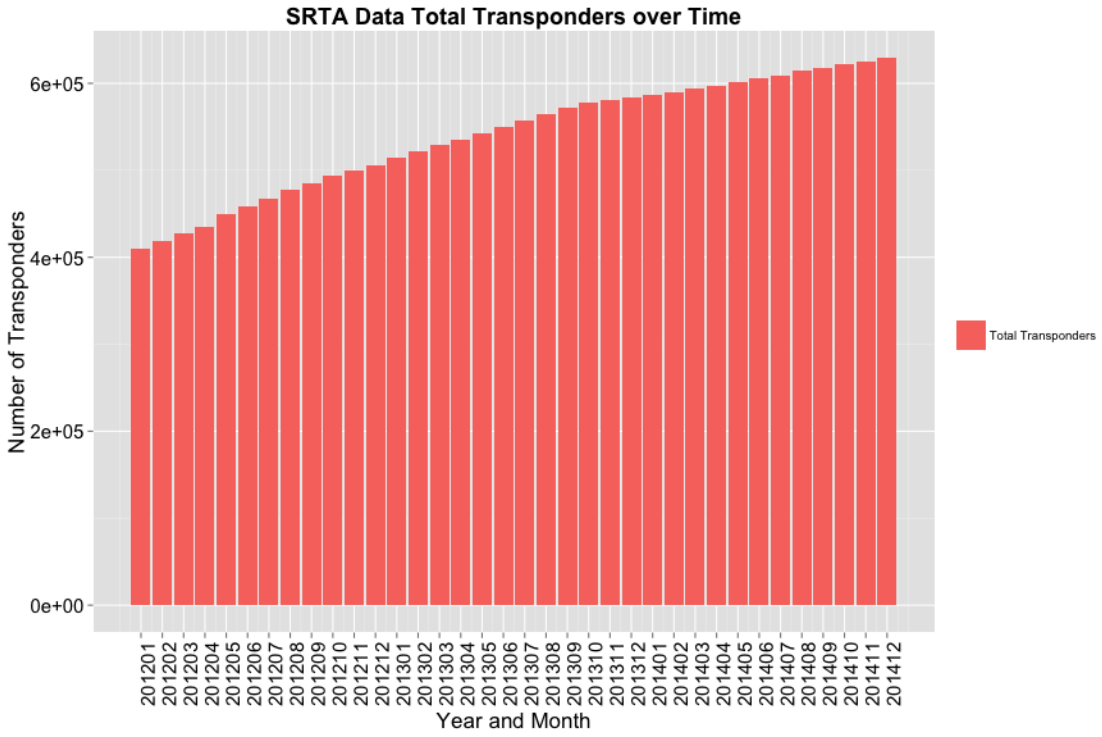


Figure 79: Total SRTA Transponders over Time

Figure 80 presents the total number of SRTA-registered transponders in Active status over three years of operation. Similar to plot of total active accounts over time, this figure shares a shape with the figure of total SRTA transponders (active and inactive) over time. In this case, the proportion of transponders in active status ranges from 72.1% to 85.5%.

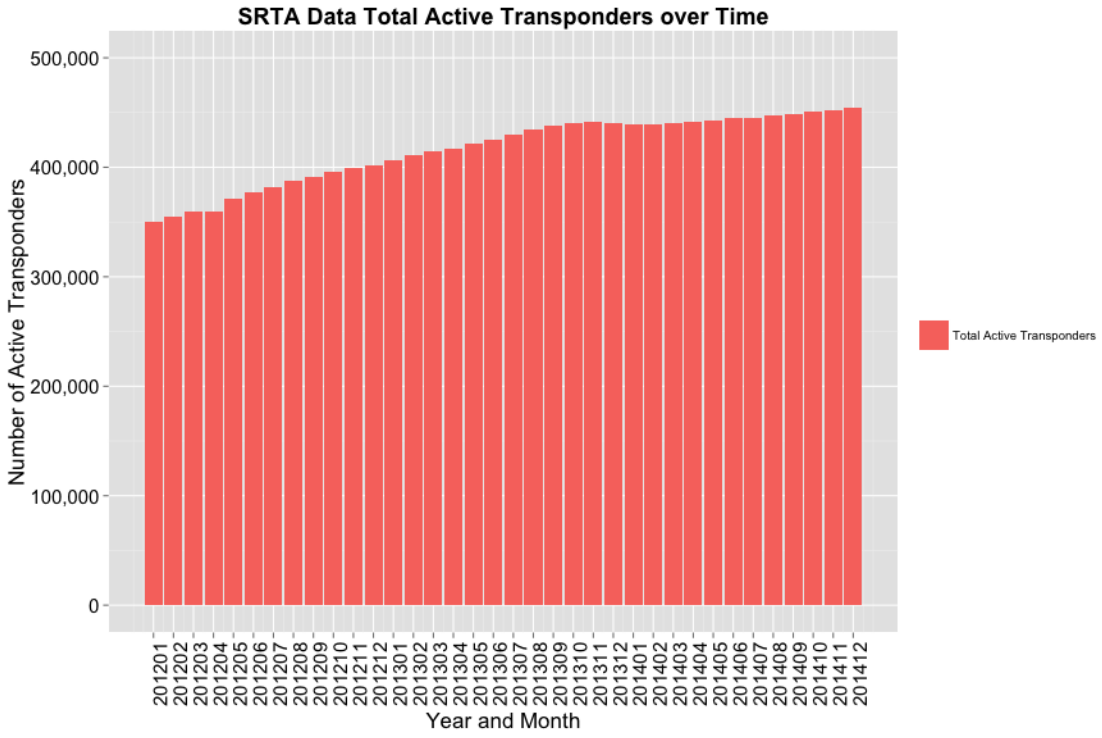


Figure 80: Total SRTA Active Transponders over Time

Figure 81 shows the numbers of newly detected and no longer detected transponders in the SRTA Account stream. The number of new transponders in the Account data each month is far lower than the numbers of new transponders detected in the vehicle detection data each month; Figure 73 showed those new transponders regularly exceeding 20,000. That figure also showed far more transponder dropouts each month as well. The registration data, as presented in the Account records, exhibit far less variability than the actual lane use data, as represented by the vehicle detection records.

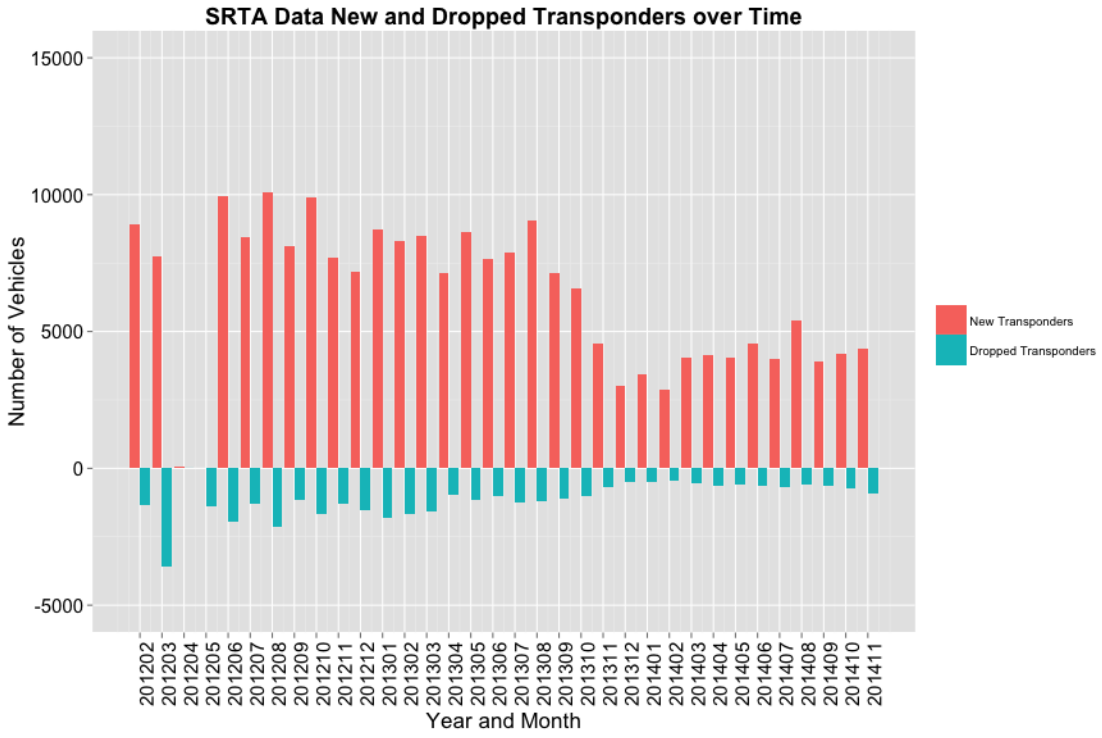


Figure 81: SRTA Data New and Dropped Transponders over Time

The potential effects of this changing sample would appear in Figure 82 below, which illustrates the GTRI registration database match rate over the scope of the study. As mentioned above, the GTRI researchers executed the registration database query on May 23, 2014. Python scripts written for this dissertation perform the matching process between the SRTA vehicle data and the GTRI registration database daily. The result is a match rate that changes every day but is very consistent month-to-month. Note again that the April, 2012 data from the Account stream was incomplete, resulting in the low match rate for that month shown here.

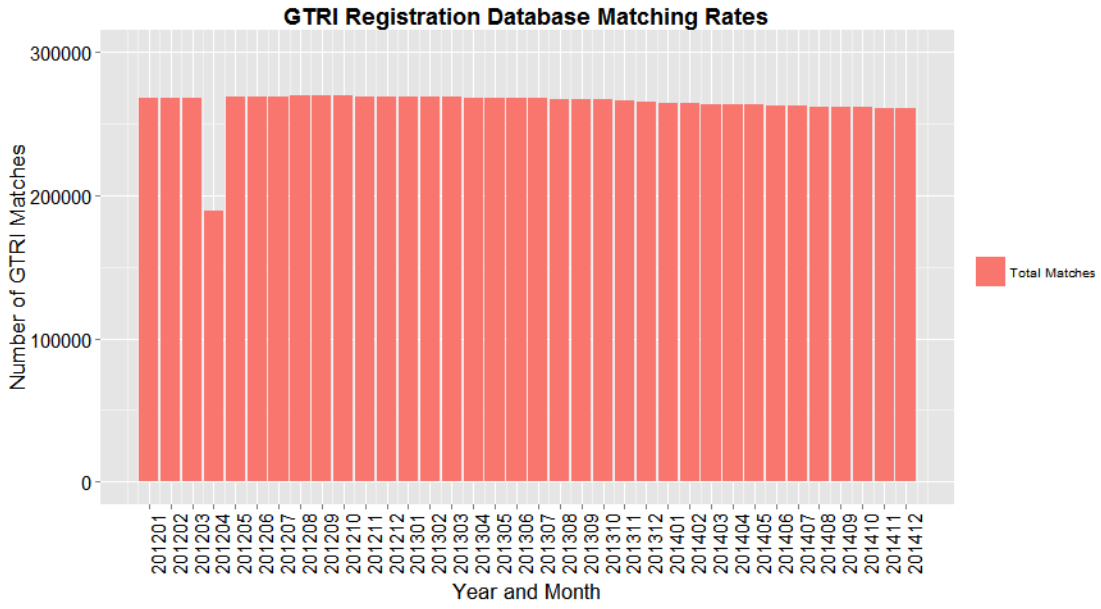


Figure 82: GTRI Match Rate over Time

Figure 83 also shows the new and dropped GTRI registration database matches each month. The number of new matches steadily decreases, to the point where very few matches are added in all of 2014. The number of dropped registration database matches also decreases in 2014, though not as drastically. The number of new and dropped matches is very small compared to the overall match rate: less than 1% of the total number of matches each month.

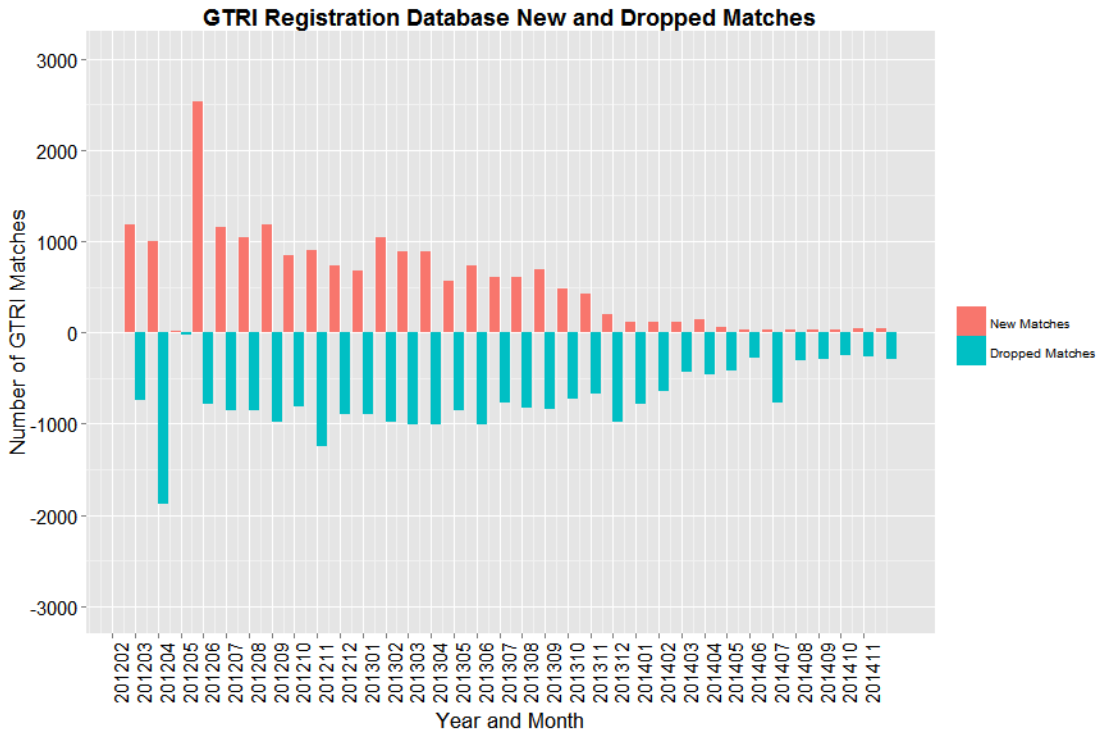


Figure 83: New and Dropped GTRI Matches over Time

Similarly, Figure 84 shows the total number of unique matches for the Epsilon households each month from January of 2012 through December of 2014. This match rate is also very consistent across the study timeframe, with an average of 46,400 households matched per month. Figure 85 illustrates the new and dropped matches each month; these were plotted separately as the difference in scale between this and Figure 84 is substantial. New matches are those whose first occurrence was in that month, while

dropped matches are those whose last occurrence was in that month. Like the GTRI registration database match rate, the number of new and dropped Epsilon households each month is multiple orders of magnitude smaller than the total number of matches. The changes in the SRTA account and lane use data appear to have little impact on the number of households in the final paired sample.

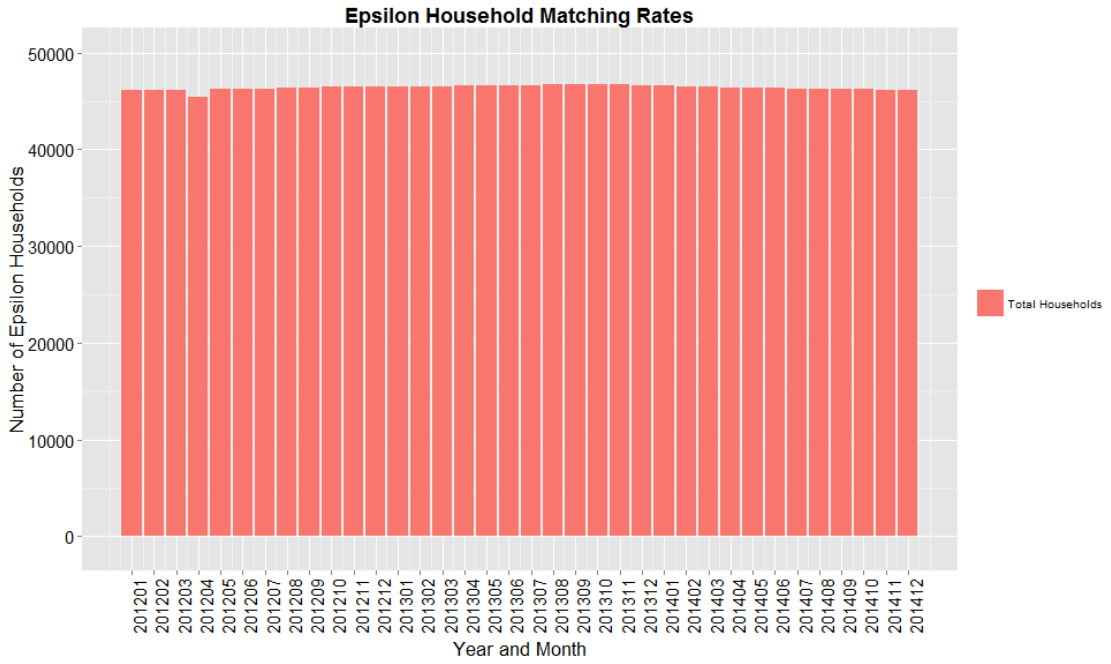


Figure 84: Epsilon Match Rate over Time

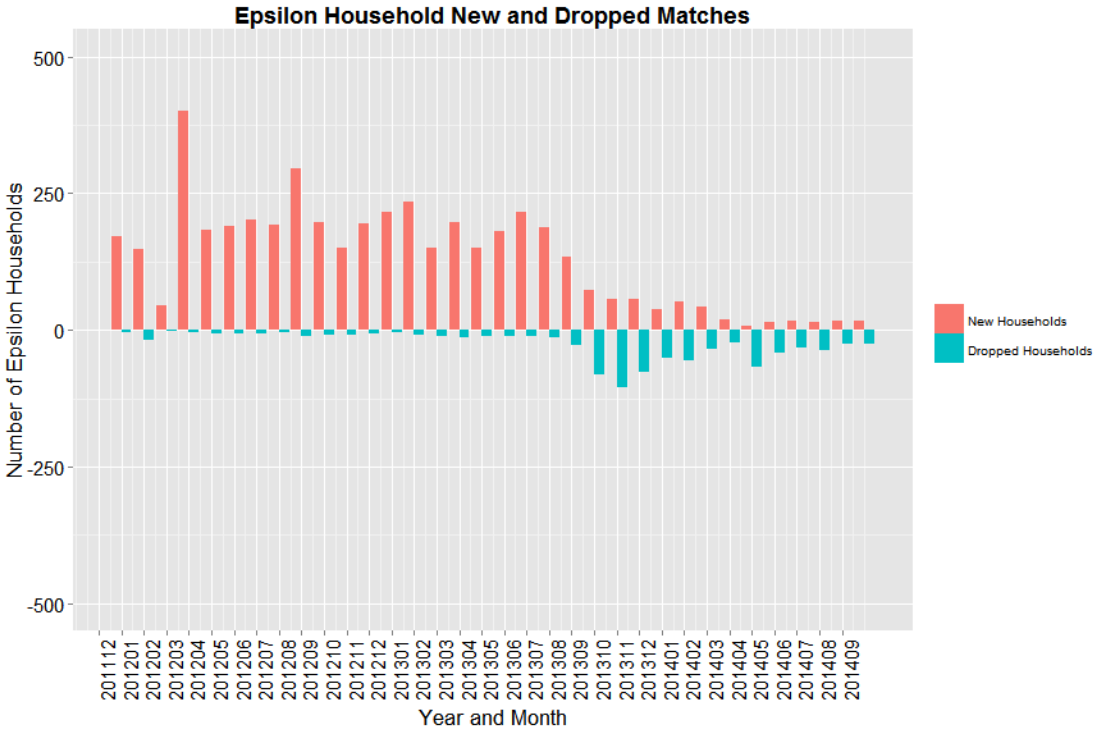


Figure 85: New and Dropped Epsilon Matches over Time

The constantly-changing sample highlights a characteristic of the SRTA and Epsilon data sets. Thousands of new transponders are registered each month, and more transponders are detected for the first time each month. The match rate at both the GTRI and Epsilon levels changes far less dramatically, however. This discrepancy indicates that though the actual composition of lane users is constantly changing, the users being studied are relatively constant. The benefit of such a result is that many users can be studied in a longitudinal fashion; very few Epsilon households drop out of the sample each month. The downside of such a data set lies in the large proportion of users who fall outside of the scope of the study, especially those that entered the system in the last six months of 2014.

Data Issues in Epsilon Demographic Data

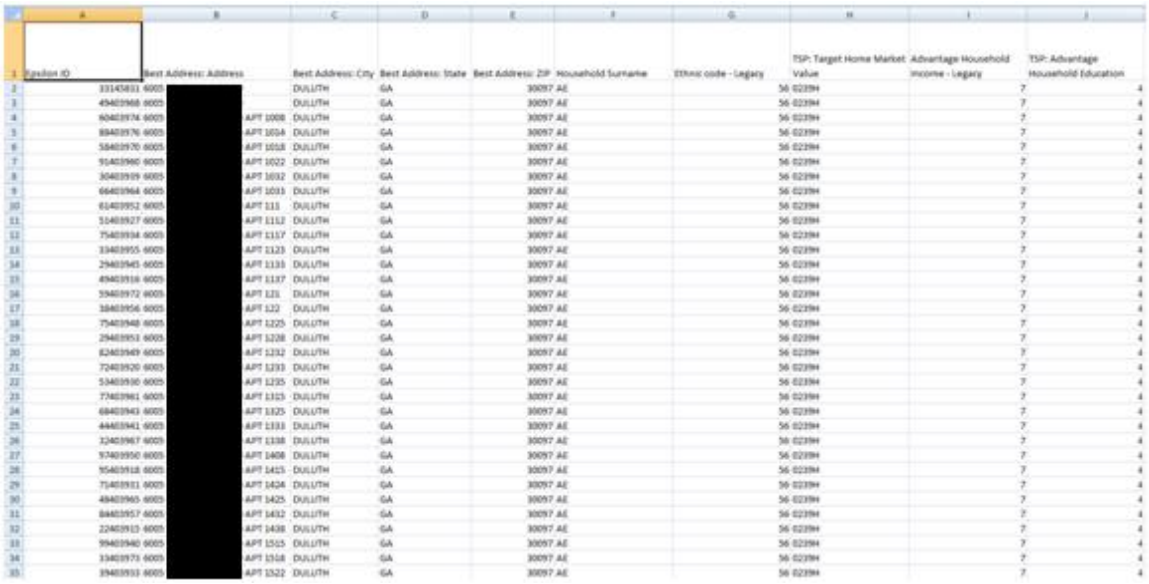
In the process of performing the data processing and treatment described here, a number of issues with the private marketing data arose were recognized as potential avenues for bias in the sample. As discussed above, the marketing data do not have complete coverage; many of the variables have imputed values for some households. The method of imputation is confidential and unknown to Georgia Tech. The timeframe of the marketing data may also not match up exactly with that of the trip data. Khoeini (2013) described how many households experience changes in different socioeconomic characteristics over the two-year period of the 2006 Commute Atlanta study. For example, 18% of the households in that study saw a change in household income, for example. As a result, some of the marketing data may be outdated relative to trip dates. These potential sources of bias may all add uncertainty to the disaggregate choice-based

analysis. Among those issues was Epsilon's initial handling of the households within an apartment complex.

Epsilon Data Multi-Family Dwelling Unit Issue

In examining the household demographic data, researchers identified several instances of names and variable values that were identical. Upon further investigation, researchers found that these records belonged to multi-family buildings such as apartment and condominium complexes. The team concluded that household records from the same multi-family dwelling units were assigned the same name and demographic data in the purchased data set, despite the records pertaining to different units within those buildings. It appeared that the query used to pull records from the Epsilon database assigned the first available value to an address, irrespective of the second address line. In short, all of the records from a given street address had the same name and variable values, regardless of the number of unique households at that address.

For example, the multi-unit building shown below in Figure 86 has 63 records for the different apartments in the Epsilon data. Each of the 63 records is listed with the same name. The figure also illustrates selected demographic data associated with each of those records; all of the various elements are identical.



	A	B	C	D	E	F	G	H	I	J
1	Epsilon ID	Best Address: Address	Best Address: City	Best Address: State	Best Address: ZIP	Household Surname	Ethnic code - Legacy	TSP: Target Home Market Value	Advantage Household Income - Legacy	TSP: Advantage Household Education
2	33345831 6005	[redacted]	DULUTH	GA	30097	AE		56	023IM	7 4
3	45403968 6005		DULUTH	GA	30097	AE		56	023IM	7 4
4	9403974 6005	APT 1008	DULUTH	GA	30097	AE		56	023IM	7 4
5	8403976 6005	APT 1014	DULUTH	GA	30097	AE		56	023IM	7 4
6	5403970 6005	APT 1018	DULUTH	GA	30097	AE		56	023IM	7 4
7	91403980 6005	APT 1022	DULUTH	GA	30097	AE		56	023IM	7 4
8	30403939 6005	APT 1032	DULUTH	GA	30097	AE		56	023IM	7 4
9	66403964 6005	APT 1033	DULUTH	GA	30097	AE		56	023IM	7 4
10	61403952 6005	APT 111	DULUTH	GA	30097	AE		56	023IM	7 4
11	51403927 6005	APT 1112	DULUTH	GA	30097	AE		56	023IM	7 4
12	75403934 6005	APT 1117	DULUTH	GA	30097	AE		56	023IM	7 4
13	13403955 6005	APT 1123	DULUTH	GA	30097	AE		56	023IM	7 4
14	29403945 6005	APT 1133	DULUTH	GA	30097	AE		56	023IM	7 4
15	49403916 6005	APT 1137	DULUTH	GA	30097	AE		56	023IM	7 4
16	59403972 6005	APT 121	DULUTH	GA	30097	AE		56	023IM	7 4
17	38403956 6005	APT 122	DULUTH	GA	30097	AE		56	023IM	7 4
18	75403948 6005	APT 1225	DULUTH	GA	30097	AE		56	023IM	7 4
19	25403951 6005	APT 1228	DULUTH	GA	30097	AE		56	023IM	7 4
20	82403949 6005	APT 1232	DULUTH	GA	30097	AE		56	023IM	7 4
21	72403920 6005	APT 1233	DULUTH	GA	30097	AE		56	023IM	7 4
22	53403930 6005	APT 1235	DULUTH	GA	30097	AE		56	023IM	7 4
23	77403981 6005	APT 1313	DULUTH	GA	30097	AE		56	023IM	7 4
24	68403943 6005	APT 1325	DULUTH	GA	30097	AE		56	023IM	7 4
25	44403941 6005	APT 1333	DULUTH	GA	30097	AE		56	023IM	7 4
26	12403967 6005	APT 1338	DULUTH	GA	30097	AE		56	023IM	7 4
27	97403950 6005	APT 1408	DULUTH	GA	30097	AE		56	023IM	7 4
28	95403918 6005	APT 1415	DULUTH	GA	30097	AE		56	023IM	7 4
29	71403931 6005	APT 1424	DULUTH	GA	30097	AE		56	023IM	7 4
30	49403965 6005	APT 1425	DULUTH	GA	30097	AE		56	023IM	7 4
31	84403957 6005	APT 1432	DULUTH	GA	30097	AE		56	023IM	7 4
32	22403913 6005	APT 1438	DULUTH	GA	30097	AE		56	023IM	7 4
33	99403940 6005	APT 1515	DULUTH	GA	30097	AE		56	023IM	7 4
34	13403973 6005	APT 1518	DULUTH	GA	30097	AE		56	023IM	7 4
35	39403933 6005	APT 1522	DULUTH	GA	30097	AE		56	023IM	7 4

Figure 86: Example of Duplicate Epsilon Data

The extent of this issue is illustrated in Table 24. Researchers used multiple criteria to identify potentially problematic Epsilon records. Those criteria involved the residence type as specified by Epsilon, the address and last name values, and the presence of apartment or suite numbers in the apartment text. The first category of records, comprising of the rows labeled "Multi-Family Dwelling Unit records," "Condo records," "Business records," "Blank records," and "Mobile Home records," included all households that were not designated by Epsilon as "Single Family Dwelling Unit" households. A total of 46,567 records fell into this category. The second group includes records that had similar address and surname values in the dataset. Researchers wrote a script that extracted both the first ten characters of the street address and the last name from each Epsilon record and then searched for instances of duplicate values. The script used the first ten characters of the street address to avoid including apartment or suite numbers, which could be different. The script identified 42,696 total records as

duplicates; within those records were 12,912 unique values. The final category of problematic records included those with 'APT' or 'SUITE' included in the address text. A search of the Epsilon dataset identified a total of 23,522 records that fit this criteria. Many records fell into multiple categories; a record may have been listed as 'Multi-family dwelling' units and also included 'apt' in the address text. In the three categories described above, the investigation identified a total of 68,180 unique records.

Table 24: Problematic Epsilon Records

Total number of records	349,134
Multi-Family Dwelling Unit records	30,931
Condo records	10,397
Business records	260
Blank records	2,760
Mobile Home records	2,219
Records containing duplicate values	42,696
Unique addresses within duplicate records	12,912
Records with second-level addresses:	
Apt	20,303
Suite	3,219
Total number of problematic records	68,180

This issue presented an immediate and obvious source of bias for the study. Researchers must either use the same demographic data for each unit within the complex, which would be incorrect, or remove the multi-unit household data from the demographic sample. That solution generated a clear source of bias by excluding households which typically have lower incomes. Table 25 below illustrates the average values of household size, education, income, and head of household age for the single-unit, multi-unit, and other home types in the marketing data set. Households categorized in the Multi-Family Dwelling Unit category have far lower average household incomes, and their average head of household age is younger too. Education levels are similar, but single-family dwelling unit households include one more member on average.

Table 25: Demographic Means by Dwelling Type

Measure	Single-Family Dwelling Units	Multi-Family Dwelling Units	Other Units
Number of records	302,567	30,931	12,876
Household Income	$64,553	$40,047	$55,298
Household Size	2.51	1.52	1.91
Household Education Level	3.04	3.02	3.21
Head of Household Age	3.80	2.96	3.57

Figure 87 illustrates the household income distributions for households in the single family dwelling unit and multi-family dwelling unit categories. As indicated by the differences in income averages and medians, the multi-family dwelling households have incomes which are more heavily concentrated towards the lower end of the spectrum. The average household income of the multi-family dwelling category is 38% lower than that of the single family unit households.

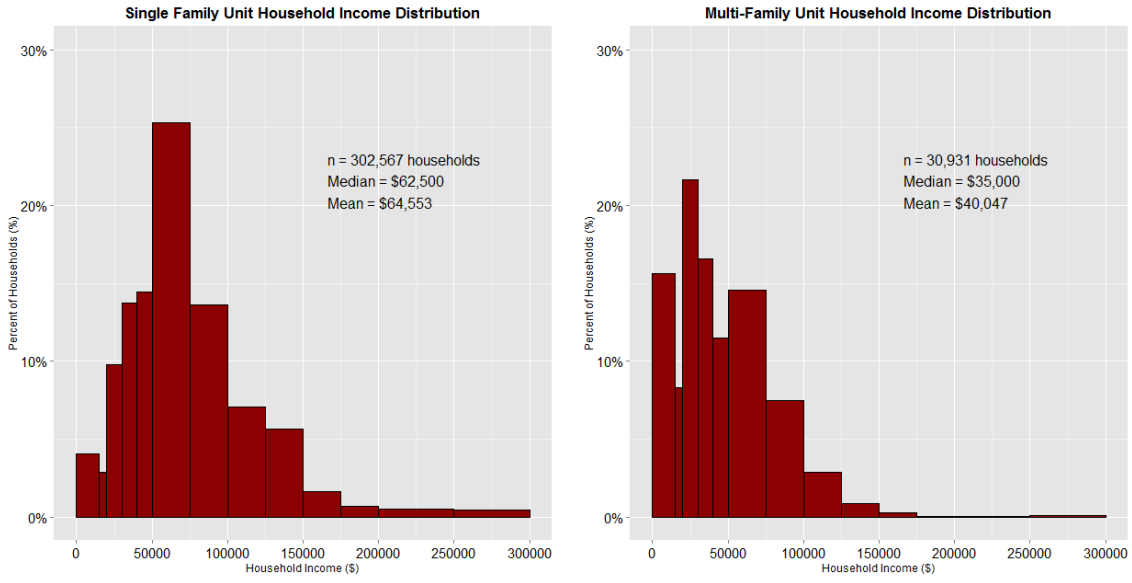


Figure 87: Household Income - Single Family and Multi-Family Units

Figure 88 shows the distributional differences in household size between the single family unit and multi-family unit households. As may be expected, households in

single family dwelling units are larger by one individual on average. The proportion of households with one individual is over 20% higher in the multi-family unit dataset versus the single family unit sample.

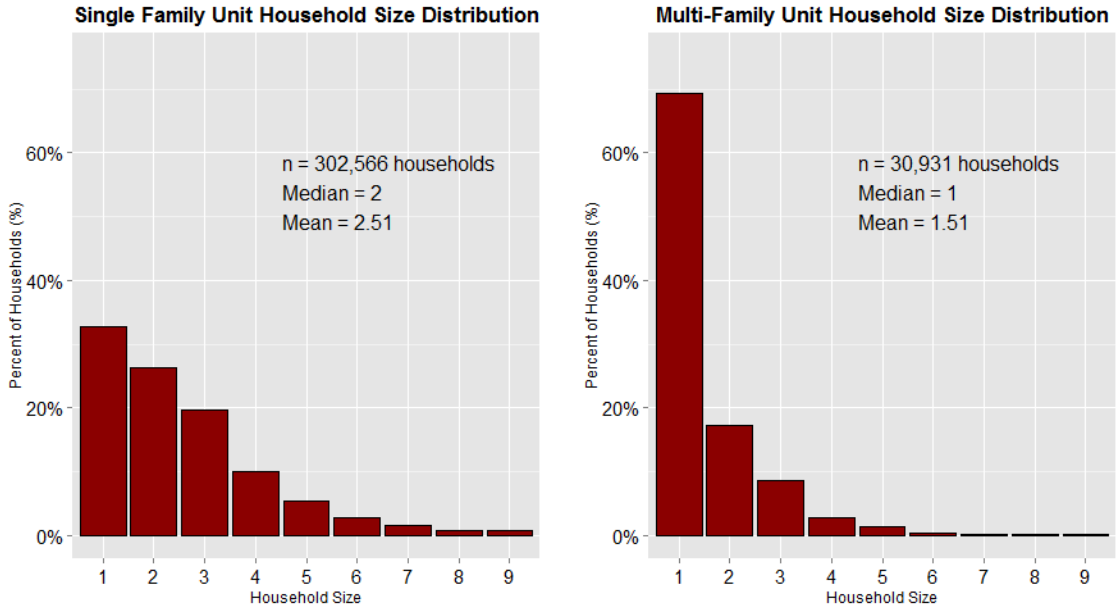


Figure 88: Household Size - Single Family and Multi-Family Units

Figure 89 illustrates the differences in household education levels for the two dwelling-unit types. Here the differences are visible but very minor; single family unit households have marginally more representation in the 'Some College' category while multi-family unit households have a very slightly higher rate of observations in the 'College' category. The 'Some High School' proportion is also slightly higher in the multi-family dwelling unit sample.

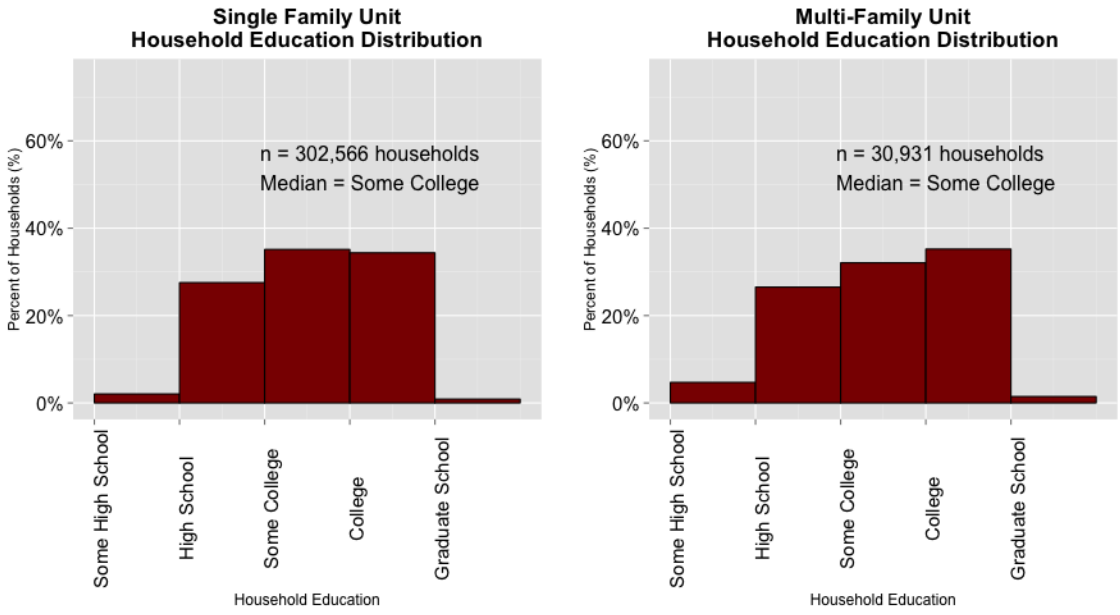


Figure 89: Household Education - Single Family and Multi-Family Units

Figure 90 shows the differing head of household age values for the single family and multi-family unit households. Here the differences are once again pronounced. Households in single-family units skew older: the proportion of households in the 25-34 age bracket is over 30% higher in the multi-family unit distribution. The results indicate that the multi-family dwelling unit households under examination here are primarily younger people with smaller families, rather than older couples that may be downsizing after retirement.

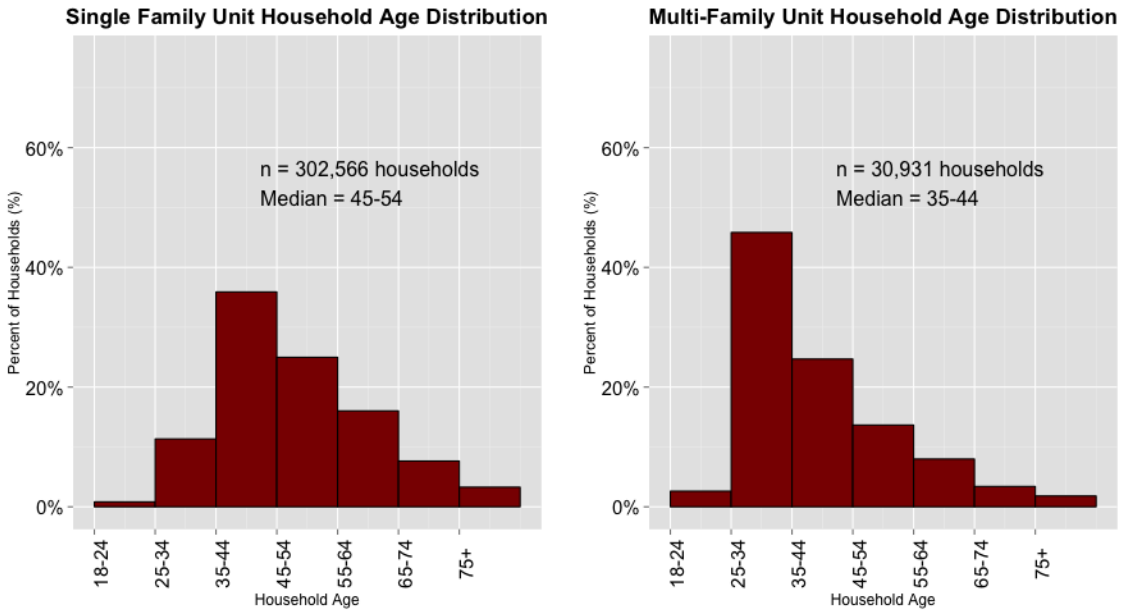


Figure 90: Head of Household Age - Single and Multi-Family Units

A summary of the differences in the single family dwelling unit and multi-family dwelling unit households can be seen below in Table 26. The income, size, and age distributions are all significantly different across the two dwelling unit types, as expected, given the charts presented above. Only the education variable cannot be said to differ, as the Mann-Whitney test could not reject the null hypothesis of distributional equality with 95% confidence. Multi-family dwelling unit households in this data set have lower household incomes, fewer members, and younger members overall.

Table 26: Differences between Single Family and Multi-Family Dwelling Data

	Single Family Units				Multi-Family Units			
	Income	Size	Education	Age	Income	Size	Education	Age
Mean	$64,553	2.51	3.04	3.80	$40,047	1.52	3.02	2.96
Median	$62,500	2	3	4	$35,000	1	3	3
Skewness	1.54	1.39	-0.20	0.56	1.77	2.47	-0.29	1.16
Kurtosis	6.79	5.11	2.04	2.86	9.18	11.01	2.18	3.86
Mann-Whitney Results	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	$p = 0.073$	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	$p = 0.073$	$p < 2.2x10^{-16}$

Revised Epsilon Demographic Data

After researchers pointed out the problems in the multi-family dwelling unit Epsilon data, the firm agreed to re-analyze the affected addresses and return new database query results to Georgia Tech. The new data set contained reprocessed records for the 68,180 previous records that researchers identified as problematic. Of those 68,180 new observations, 19,344 remained the same as the old records while 48,846 were modified. This section provides an investigation of the extent of the differences. Figure 91 shows the distributions of annual household income for the households that were identified as having problematic data in the original data set. The re-processed results for those households that were returned with different values are shown on the right hand side. The mean and median values of the two distributions are close; the most notable difference appears to be fewer households in the lower income range in the re-processed records. Note that the difference in sample size values for the original and re-processed data reflect missing values in the original data that were populated in the newer dataset.

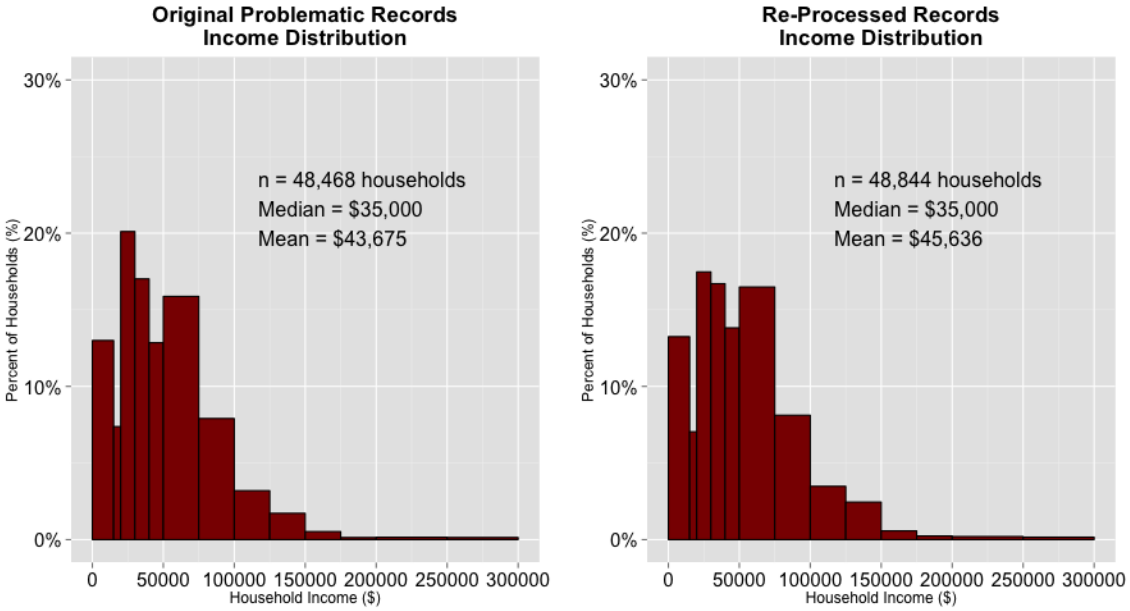


Figure 91: Household Income Distribution for Old and Re-Processed Data

Figure 92 shows the differences in the household size distribution after re-processing. Here the biggest change appears in the '1' category: the re-processed data has noticeably fewer households with that size and thus has a higher average value with a flatter, more right-shifted distribution.

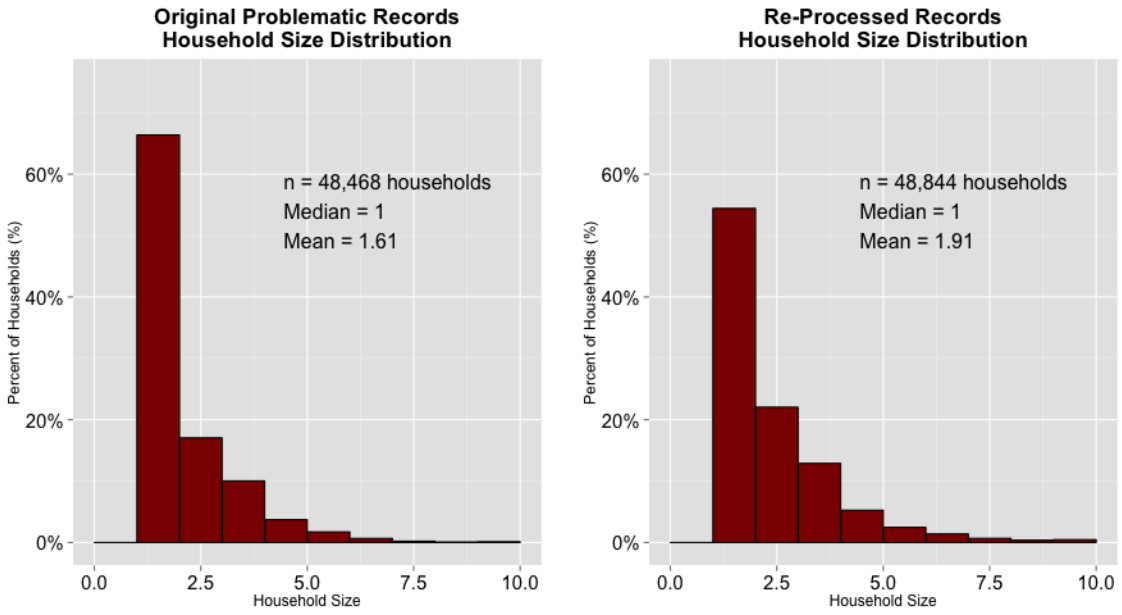


Figure 92: Household Size Distribution for Old and Re-Processed Data

Figure 93 shows the differences in education level for the affected households. Here the two distributions are very similar, with no notable (or visible) differences between them. Of the four factors examined here, this is the most consistent.

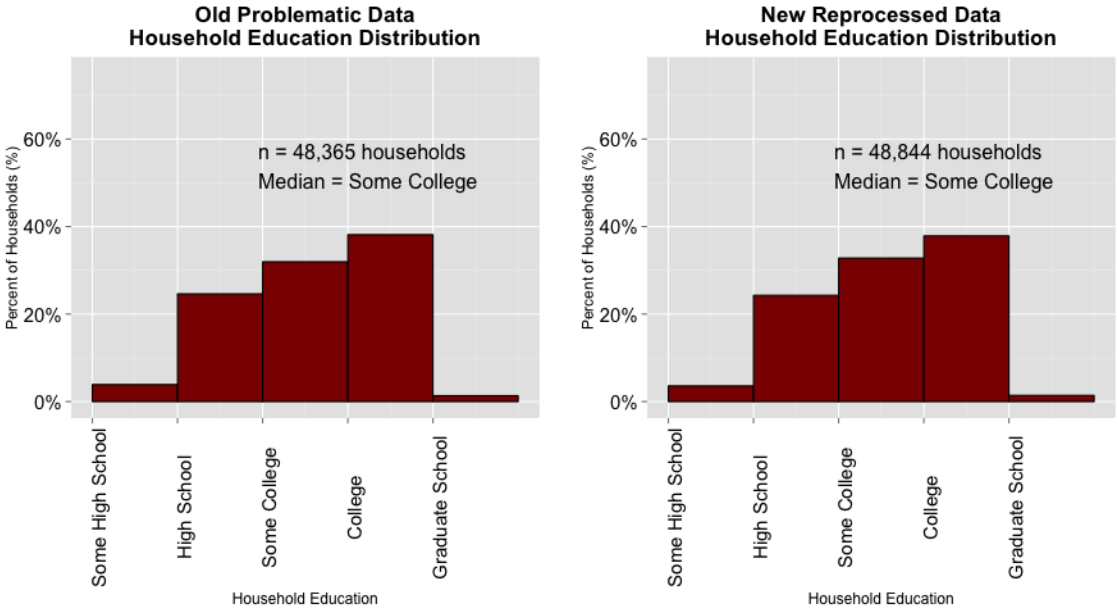


Figure 93: Household Education Distribution for Old and Re-Processed Data

Figure 94 presents the differences in the head of household age distributions among the old and re-processed Epsilon demographic data. Again, the median head of household age value remains the same for the two data sets, while the overall distribution flattens out and shifts to the right due to fewer households in the 25-34 age range category.

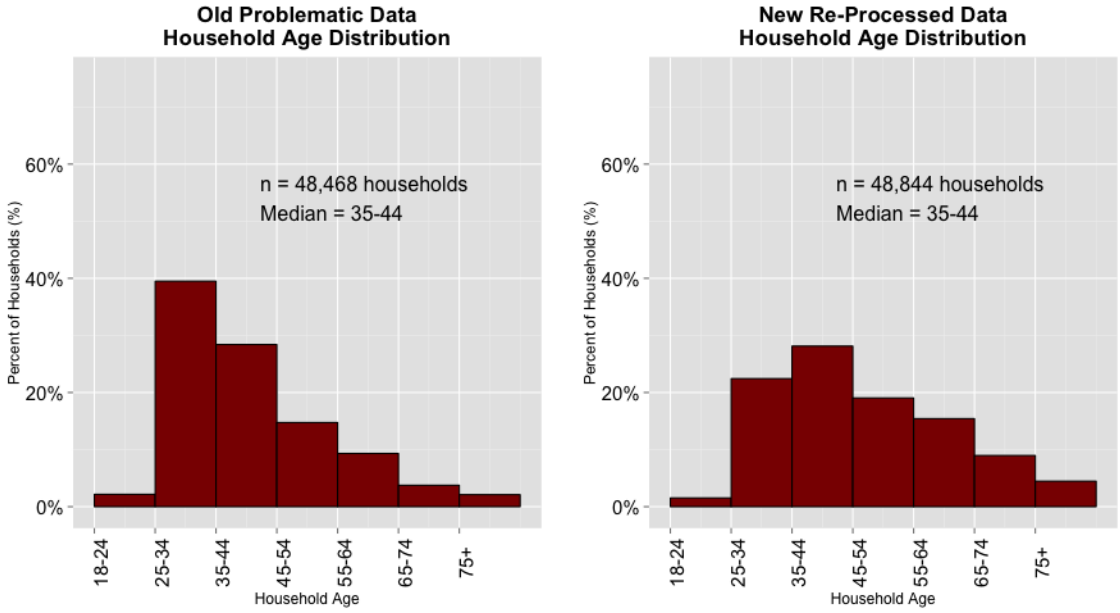


Figure 94: Head of Household Age Distribution for Old and Re-Processed Data

Overall, the re-processed data show some significant differences for the nearly 49,000 households that were affected. Table 27 summarizes these differences. The households in the re-processed data have slightly higher annual household incomes: while the mean difference is less than $2,000, that difference represents a 4.5% increase over the original income average. This is represented visually by a flatter, slightly right-shifted distribution for the re-processed data; the lower skewness and kurtosis values confirm these differences. Similarly, the re-processed households are larger by 18.6% on average. The re-processed distribution displays less of a peak at the value of one, and is

thus also flatter and shifted more to the right than the original problematic data distribution. The mean value for the head of household age increased by 19.4%, primarily by removing households from the 25-34 category. Of the four factors examined here, only household education saw no significant changes after re-processing This is shown in the high degrees of similarity in the mean values and the skewness and kurtosis results. This category was the only one in which the Mann-Whitney distributional comparison test could not reject the null hypothesis of equal distributions. Overall, the reprocessed data appears to have addressed a bias towards small, younger households with slightly lower annual incomes that was present in the original dataset. Note that the initial analyses presented in Chapter 8 and the beginning of Chapter 10 use the original data, while the later analyses use the corrected data.

Table 27: Summary of Differences Between Old and Re-Processed Data

	Original Problematic Data				New Re-Processed Data			
	Income	Size	Education	Age	Income	Size	Education	Age
Mean	$43,675	1.61	3.08	3.09	$45,636	1.91	3.09	3.69
Median	$35,000	1	3	3	$35,000	1	3	3
Skewness	1.87	2.31	-0.37	1.03	1.82	2.11	-0.36	0.51
Kurtosis	9.16	9.77	2.21	3.56	8.53	8.50	2.23	2.42
Mann-Whitney Results	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	$p = 0.3514$	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	$p < 2.2x10^{-16}$	0.3514	$p < 2.2x10^{-16}$

Quality of SRTA Vehicle Detection Data

Another foundational element of the analytical dataset that required quality assurance was the individual RFID vehicle detection data provided by SRTA. Previous examinations of the data revealed potential issues in detection reporting, particularly in the timestamps associated with vehicle detections. The SRTA lane use data also suffered from transmission issues that interrupted the data streams and individual gantry-level reporting issues stemming from faulty hardware or other causes that resulted in abnormally low detection counts. This section will provide an overview of these complications.

Mistimed Gantry Detections

The first of these issues was the occurrence of misreported gantry detection times. The table below shows an example of the detections of a single transponder over a six-minute period. The bolded rows illustrate an instance of a detection that appears to have been reported at an incorrect time. The detection at the fifth Old Peachtree Southbound gantry is reported after the detection at the third gantry, though it is physically located immediately after the sixth Southbound gantry. This would result in the trip-building script splitting the detections up into two different trips, despite the proximity of the detection times and the otherwise logical spatial progression of the detections.

As a result, the trip-building script was modified to allow these misdetections while keeping track of the number that occur within each trip. This changed the number of trips that were generated by the script: for an example day (February 15th, 2012), the total number of trips was reduced from 32,762 to 32,608. Out of those 32,608 trips, 158 included misdetections. One of those trips had two misdetections and the remaining 157 had one misdetection. Incorporating the misdetections into the trips in this way also changed the speed characteristics of the resulting

trips. For the February 15, 2012 example, the algorithm that broke up trips with misdetections yielded 28 trips with speeds over 100mph; two of these trips were estimated to have speeds of nearly 200mph. After incorporating the misdetections so that the trips were not broken up, there were nine trips with speeds over 100mph. The maximum speed in this new set was 114mph. This change appeared to reduce the number of unreasonably high speeds that were the result of misreported detections.

Table 28: Example of Misreported Detection

LaneID	TransactionDateTime	Direction	Gantry
170400	2/16/2012 13:20:24	SB	OP09
170390	2/16/2012 13:20:49	SB	OP08
170380	2/16/2012 13:21:10	SB	OP07
170370	2/16/2012 13:21:30	SB	OP06
170348	2/16/2012 13:22:06	SB	OP04
170338	**2/16/2012 13:22:25**	**SB**	**OP03**
170359	**2/16/2012 13:22:34**	**SB**	**OP05**
170328	**2/16/2012 13:22:44**	**SB**	**OP02**
170295	2/16/2012 13:23:52	SB	PH07
170281	2/16/2012 13:24:09	SB	PH06
170267	2/16/2012 13:24:33	SB	PH05
170256	2/16/2012 13:24:58	SB	PH04
170244	2/16/2012 13:25:22	SB	PH03
170231	2/16/2012 13:25:55	SB	PH02
170218	2/16/2012 13:26:17	SB	PH01

Figure 95 presents the counts of mistimed detections in the constructed trip set per month from January, 2012 through December, 2014. After starting from a very large baseline in the first month, the misdetection counts drop to more reasonable levels by March, 2012. Other than two relative spikes in August, 2012 and October, 2014, the timing issues in the reporting system appear to have been addressed by SRTA or their contractors.

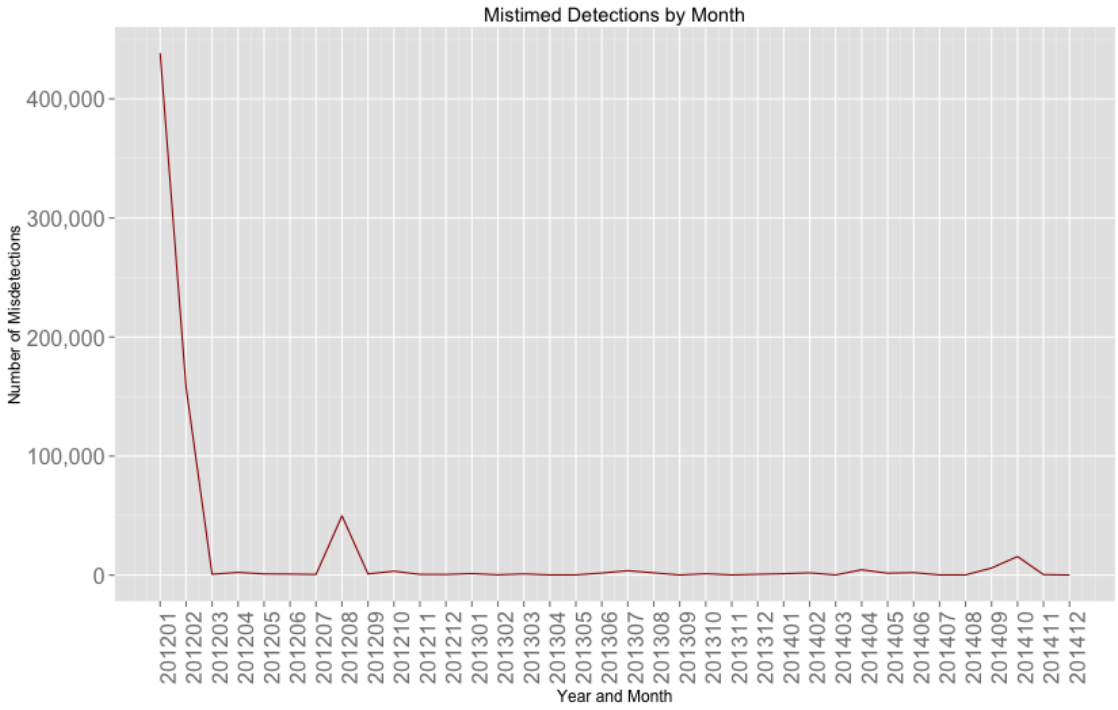


Figure 95: Mistimed Detections by Month

Interruptions in Data Transmission

Another significant issue occurred as the result of gaps in the various data streams. These gaps occurred for two primary reasons: an outage in the server link connecting Georgia Tech with SRTA/ETC, or an error in the reporting system. Instances of the first type of gap, in which Georgia Tech stopped receiving data from SRTA, are outlined below in Table 29 though Table 31 and occurred over the course of the facility lifespan. The first table lists the dates of missing or corrupted Account data. These errors were largely concentrated in the first five months of

2012; that time frame covers 69 of the 72 missing days' worth of data. The 72 total missing days represent 6.6% of all days in 2012-2014. For data processing steps that require daily account data, such as those that pair the daily active transponders with Epsilon marketing data, the scripts find the first or most recent valid account file in that month. For example, a script that looks for the account file for March 7, 2012, will instead use the records from March 6, 2012. In the case of May, 2012, in which the first three days of data are missing, the scripts identify the 4th as the first available date and use those records for May 1-3.

Table 29: Gaps in SRTA Account Data Transmission

Start Date	End Date	Days Missing
1/6/2012	1/7/2012	2
1/11/2012	1/11/2012	1
1/14/2012	1/14/2012	1
1/21/2012	1/22/2012	2
1/24/2012	1/25/2012	2
1/27/2012	1/28/2012	2
2/1/2012	2/1/2012	1
2/3/2012	2/5/2012	3
2/7/2012	2/10/2012	4
2/18/2012	2/18/2012	1
2/25/2012	2/25/2012	1
3/7/2012	3/7/2012	1
3/11/2012	3/11/2012	1
3/13/2012	3/13/2012	1
3/15/2012	3/17/2012	3
3/21/2012	3/24/2012	4
3/27/2012	5/4/2012	39
2/13/2014	2/14/2014	2
2/28/2014	2/28/2014	1
Total		72

Table 30 provides the time frames for the gaps in the remote traffic microwave sensor (RTMS) data stream used to collect vehicle counts and speeds. Data from this real-time feed cannot be recovered in the way that other files can. These gaps are less of an issue here, because this dissertation does not use the RTMS data stream. The 136 missing days represent 12.4% of the three year timespan from 2012-2014.

Table 30: Gaps in SRTA RTMS Data Transmission

Start Date	End Date	Days Missing
3/18/2012	4/13/2012	27
6/4/2012	6/7/2012	2
2/16/2013	4/18/2013	31
6/4/2013	8/9/2013	67
9/30/2013	10/8/2013	9
Total		136

Only eleven days' worth of individual vehicle detection data were lost, as shown in Table 31. The majority of these occurred in 2013. Losses in this stream are more disruptive to the analysis, as it forms the basis of the constructed trip set and many of the operational data sets to which those trips are joined, such as travel speed averages and transponder counts. Unlike the RTMS outages, the missing data in the Vehicle detection stream occurs on a day-by-day basis. Because the feed is not real time, any gaps in the transmission can be rectified by recovering the detection data once the connection has been restored. The remaining losses occur due to empty or corrupted files rather than connection errors. The 11 missing days represent just 1.0% of the 1,096 days in the three years of analysis.

Table 31: Gaps in SRTA Vehicle Data Transmission

Start Date	End Date	Days Missing
10/6/2012	10/6/2012	1
1/27/2013	1/27/2013	1
2/24/2013	2/24/2013	1
4/17/2013	4/17/2013	1
6/5/2013	6/5/2013	1
8/17/2013	8/17/2013	1
9/26/2013	9/26/2013	1
10/8/2013	10/8/2013	1
10/22/2013	10/22/2013	1
1/27/2014	1/27/2014	1
2/28/2014	2/28/2014	1
Total		11

The missing data in the Express Lane Trip summary stream are listed in Table 32. These records are fairly evenly distributed through the three years under examination. This stream is primarily used to identify which HOT trips were taken in Toll mode versus Carpool mode; other details about the trips themselves are replicated in constructed trip set which is derived from the individual vehicle detections. The 18 missing days constitute 1.6% of the total days in the three year time frame.

Table 32: Gaps in SRTA Trip Data Transmission

Start Date	End Date	Days Missing
9/21/2012	9/23/2012	3
10/28/2012	10/28/2012	1
1/27/2013	1/27/2013	1
2/24/2013	2/24/2013	1
4/17/2013	4/17/2013	1
8/17/2013	8/17/2013	1
10/8/2013	10/8/2013	1
10/22/2013	10/22/2013	1
1/27/2014	1/27/2014	1
2/13/2014	2/14/2014	2
2/28/2014	2/28/2014	1
5/11/2014	5/14/2014	4
Total		18

The overall impact of the missing data is slight; the most important stream, the Vehicle detections, has 99% of the study days represented in the data. The Trip summary stream includes over 98%. While the missing data rate for the Account stream is higher, at 6.6%, those data can more readily be substituted for with neighboring files. The most extreme case of missing data occurs in the RTMS feed, which is not used in any analysis presented here.

The second type of issue occurred primarily at the beginning of the facility operations. These reporting errors, and their durations, were as follows:

1. Between the opening of the facility on October 1, 2011 and January 6, 2012, the General Purpose lane vehicle detectors were offline. No vehicle detections were reported in the GP lanes until January 6, 2012.
2. Until January 29, 2012, the Express Lane system reported no southbound trips originating at the Pleasant Hill segment of the corridor or those that start and end in the Old Peachtree segment. These include trips ending at the end of the Pleasant Hill segment, as well as those ending at Jimmy Carter Boulevard southbound,

Issues with Vehicle Detection Gantries

In addition to identifying times in which SRTA Express Lane use data was missing or corrupted, the author investigated aberrations in the reported data itself. This section looks at the individual RFID detection stream provided by SRTA to investigate potential problems in the reporting hardware or software. A python script was employed to examine each day's worth of detection data from 2012 through 2014 and counted the daily detections at each of the 35 HOT gantries. Figure 96 shows the resulting detection counts at each gantry for the duration of 2013; the plots are separated by the corridor segment. The x-axis represents the day of the year (1-365). Within

a given corridor segment, there may potentially be great variation in the typical number of detections recorded by individual gantries. The Jimmy Carter Boulevard portion of Figure 96 demonstrates this, as one gantry consistently exceeds 10,000 detections per day while two others report less than 5,000 per day. This is also reflected in the measure of dispersion: while the average number of detections per HOT gantry per day in the 2013 data is 7435.5, the standard deviation is 4858.5.

In each case, there are six dates in 2013 on which each segment reports virtually no detections. Those six dates are the same for each corridor section and for all of the individual gantries within those sections. Within those six days, the average number of detection counts across all thirty-five gantries is 7.0 per gantry. For the remainder of the year, the average number of detection counts across all gantries is 7558.5 per gantry. The scope of the issue in the 2013 data, in terms of the proportion of affected gantries, indicates a system-wide problem rather than a gantry-specific problem.

Table 33 lists the dates over all three years of study on which the Express Lane gantries reported fewer than 100 detections, along with the number of gantries for which this occurred and the average detection count at those gantries. This list of dates includes two for which the detection count issue is not systematic but rather isolated to a small subset of one or two gantries.

Table 33: Dates of Low Express Lane Gantry Detections

Date	Number of Affected Gantries	Average Detection Count
10/13/2012	2	26.50
10/27/2012	32	13.25
2/23/2013	34	9.35
4/16/2013	34	4.09
8/16/2013	35	7.43
9/25/2013	35	7.83
10/7/2013	35	10.06
10/21/2013	35	3.34
1/26/2014	35	7.66
1/29/2014	35	50.74
2/12/2014	34	2.29
2/27/2014	35	5.06
7/15/2014	1	1.00

The thirteen days represented in Table 33 constitute 1.2% of the 1,096 days between 2012 and 2014. The impact of these days in which abnormally low numbers of detections are reported is that those days are essentially removed from the analysis. Without sufficient detections, the processing scripts cannot construct vehicle trips. The 100-detection criteria was selected to identify and isolate the six problematic dates in Figure 96. Expanding that criteria to a maximum of 500 detections changes the number of affected days to a total of twenty-five, representing 2.3% of the total number of days examined. The final analyses used the 100-detection criteria.

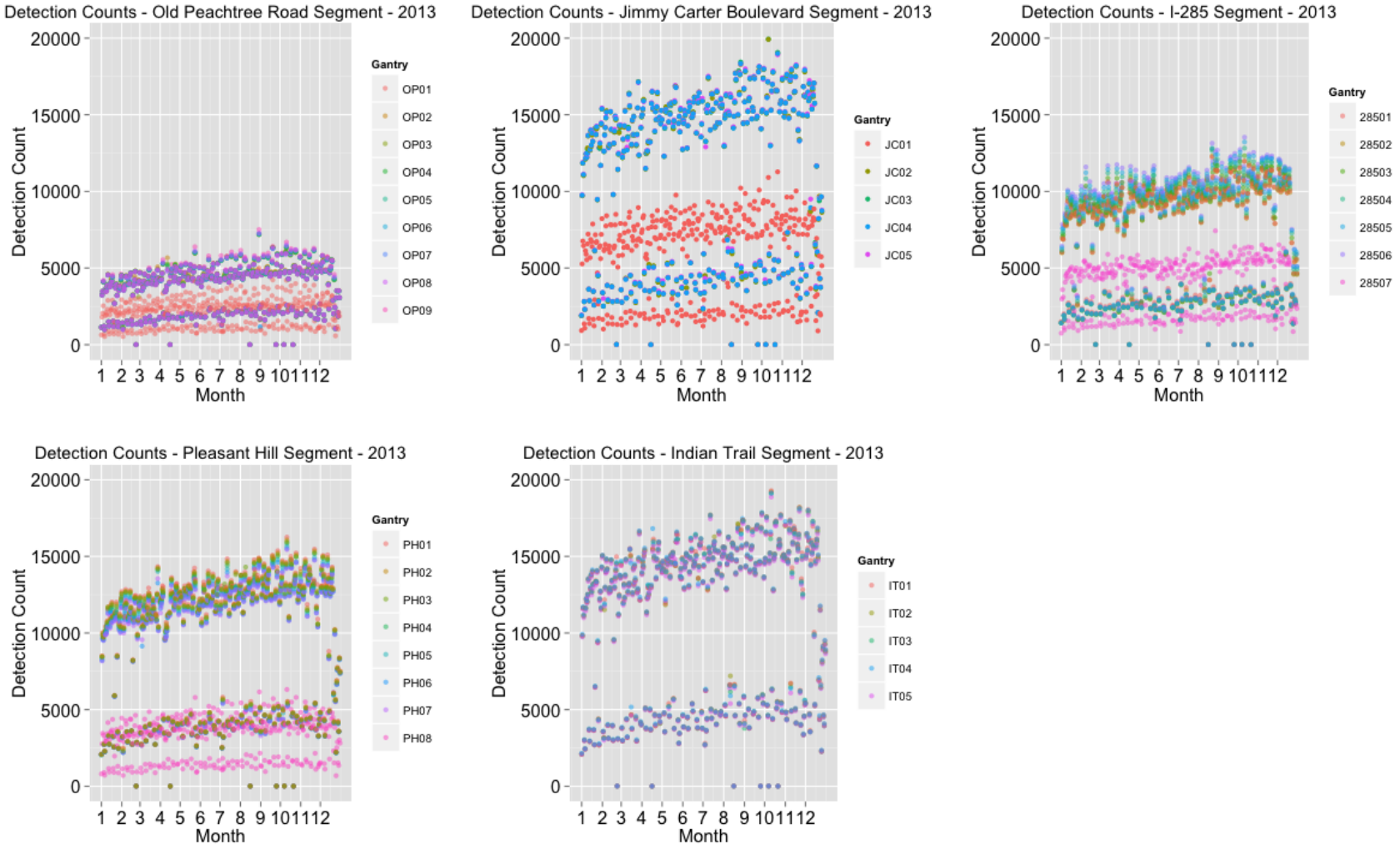


Figure 96: HOT Detection Counts by Gantry - 2013

Chapter Summary

This chapter presented an overview of the issues and complications involving the various data sets used in this dissertation. These complications involved the structure of the data, the stability of the sample, and the quality of the data themselves. The structure of the SRTA account data limited the scope of the user population that could be included in this dissertation. The need to join vehicle data to transponder data restricted the sample to those accounts with one of each, though this restriction was loosened by finding accounts whose vehicles were all registered at one address. The Express Lane Trip summary stream was able to improve the size of the sample but also suffered from many-to-many relationships that removed users from the study population. This trip stream included transponders associated with multiple plates, plates associated with multiple transponders, and records with blank data in one or both of those fields. Furthermore, these instances of transponders associated with multiple plates, or vice versa, also overlapped chronologically, narrowing and complicating the pairing process. The resulting sample of transponders and households was smaller than the original population, but still included tens of thousands of each.

Further complicating the pairing process was the longitudinal nature of the SRTA lane use data. Account records changed daily: a transponder that was active one day may have been inactive the next. Similarly, each month of lane use data included thousands of transponders that were detected for the first time. The Epsilon demographic data, however, were cross-sectional, representing household characteristics from a single point in time. The resulting rate of matches between the SRTA lane use data, the registration

database, and Epsilon demographic data was very consistent. Though the users of the lane changed constantly, the sample under examination did not.

Aside from the complications with the structure of the data, the contents of the various data sets had issues as well. The Epsilon demographic data included problematic records for apartment dwellers and other households in multi-family units, though this issue was addressed for the majority of affected households. The corrected data from Epsilon replaced the previously problematic data in the analytical file. The SRTA lane use data suffered from interruptions in the data streams that eliminated specific dates from the analyses, but this impact was small. Errors in gantry detection timing complicated the trip construction process, especially towards the beginning of the study period. Again, this had a small impact on the overall data set. The gantries were also affected by days of systematic or individual errors that reduced the numbers of reported detections to almost nothing. These days also fell outside of the scope of the analysis due to the lack of data. Some of the issues outlined here could be addressed, either through workarounds or by revising the affected data. Others could not be addressed, and thus limited the scope of the study. The primary goal of the data quality investigation was to illustrate the various ways the sample was affected due to issues in the data sources. What remains to be investigated is the overall impact of these issues on final sample. Though data quality issues narrowed the sample that was available for analysis, a sizeable number of transponders and households remain. The next chapter will investigate the potential bias in the sample that results from these data quality issues.

CHAPTER 7

POTENTIAL SAMPLE BIAS IN PAIRED VEHICLE ACTIVITY AND MARKETING DATA

Initial analyses of the pairing of SRTA vehicle activity data and Epsilon marketing data set generated unexpected results. The HOT lane use behavior among users in the lower, medium, and higher income segments was very similar (for this analysis, lower income households were defined as those with $50,000 or less in annual income, medium income households had $50,000 to $100,000, and higher income households had over $100,000 in annual income). The rates at which these groups used the HOT lanes relative to the GP lanes in the dataset, specifically trips from 2013, exhibited a 3.2% difference between the higher income and medium segments, and a 3.9% difference between the higher and lower income segments. In both cases, the higher income segment had the higher rate of use. With the limitations of that dataset in mind (no trips across both lane types, only 11% of the transponder population represented, etc.), Table 34 is reprinted here from that research to illustrate the similarities in Express Lane use rates among the different income segments (Sheikh, 2015).

Table 34: 2013 Trip Characteristics by Income Segment

	Full Dataset	Lower Income	Medium Income	Higher Income
Households Analyzed	28,953	7,959	12,592	8,402
% of Households by Income	100	27.5%	43.5%	29.0%
Total Trips Monitored	1,304,079	393,069	600,696	310,314
HOT Trips	282,616	80,340	126,745	75,531
GP Trips	1,021,463	312,729	473,951	234,783
% of Total Trips by Income		30.1%	46.1%	23.8%
% of HOT Trips by Income		28.4%	44.9%	26.7%
% of GP Trips by Income		30.6%	46.4%	23.0%
% of Trips in HOT Lane		20.4%	21.1%	24.3%
% of Trips in GP Lanes		79.6%	78.9%	75.7%
Average Trip Speed (mph)	52.3	52.1	52.4	52.6

These results did not conform with research reported in similar contemporary studies in other cities and for other HOT lane facilities, which identify household income as a major, significant factor in toll lane decision making (Li, 2001; Burris, 2006). Because of this discrepancy, this dissertation includes an investigation into potential bias in the paired lane use and demographic dataset. This chapter outlines these potential areas of bias in the sample, beginning with a look at the trip-taking behavior of users at each stage in the data pairing process. The next section examines the rate of dropouts by frequency of corridor use in the pairing process to investigate whether there is a relationship between trip frequency and pairing success. After that comes a look at the commutershed restriction employed in the analysis, followed by a comparison of available Census Bureau data at each stage in the process. The chapter then discusses issues with the structure of the Account data structure in the SRTA lane use data, and finally provides an overview of the data loss at each stage in the process.

Cumulative Trip Distributions by Sampling Level

The first step in examining the potential bias in the demographic sample was creating cumulative trip distributions for the different levels of data pairing and data loss.

These levels include the full set of all detected transponders, the transponders that could be paired with address data from the vehicle registration database, and finally transponders that had matching demographic data.

Figure 97 presents cumulative corridor trip count distributions based on the constructed trip dataset from 2013. The transponders are ranked on the x-axis by the total number of trips they took over that timeframe. The y-axis represents the share of the total trips taken. For example, in the topmost chart, examining all transponders regardless of matching status, the top 10% of corridor users (identified on the x-axis) collectively took 69.69% of the total corridor trips in 2013 (identified on the y-axis). At the median, half of the users (50^{th} percentile of users) made 97.10% of all 2013 trips. This figure includes all of the transponders in the data set; no pairing or narrowing has occurred yet. Similarly, it includes all transponder-equipped trips; not just toll lane trips.

The second chart in Figure 97 provides the same distribution for address-matched transponders. These are the transponders for which a GTRI registration database pairing could be made. This is the first of two steps in pairing the SRTA vehicle activity data with the Epsilon marketing demographic data, and it involves a narrowing of the sample from over 400,000 transponders to 172,357 transponders. Whereas in the unpaired chart, the top 10% of users had taken 69.69% of trips, here the top 10% of users took 56.75% of trips. At the median, the 50^{th} percentile of unpaired users took 97.10% of trips. After GTRI registration database matching, the 50^{th} percentile of users took 94.96% of trips.

The final chart in Figure 97 shows the ranked cumulative trip distribution within the activity-demographic matched dataset. The transponder population has narrowed to 76,051, or 18.02% of the original set of transponders. The cumulative distribution curve

has flattened once again: here the top 10% of activity-demographic matched users took 44.13% of trips, and the 50th percentile of Epsilon-matched users accounted for 93.36% of trips.

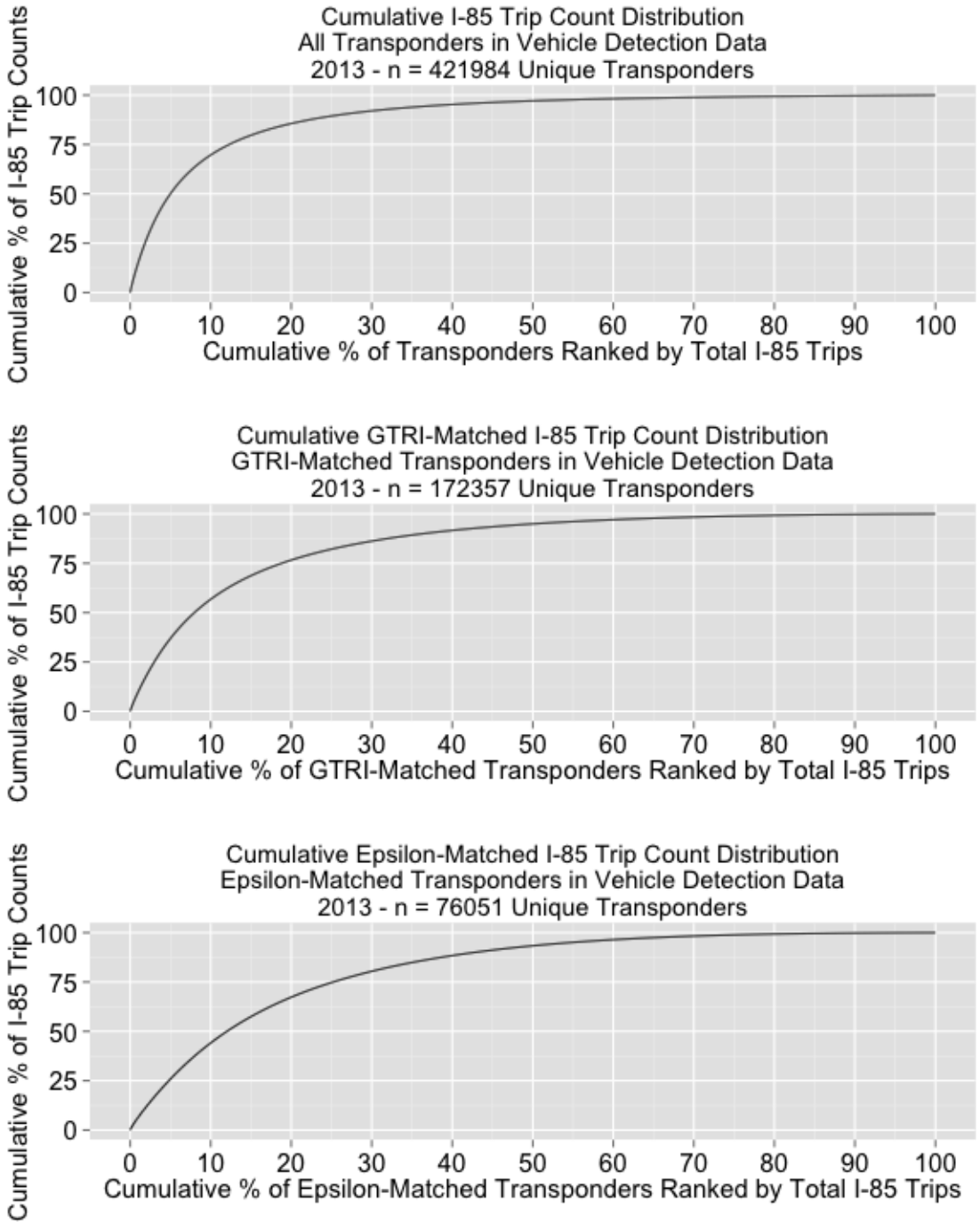


Figure 97: Trip Count Distribution by Pairing Level

Figure 98 again shows the cumulative trip count distributions at the unpaired, GTRI-matched, and demographic-matched levels. These plots restrict the trips to Express Lane trips only. As in the previous figures, each step in the pairing process creates a flattening of the cumulative distribution curve. At the unpaired level, the top 10% of users took 85.88% of all of the Express Lane trips in 2013. By the time the distribution reaches the median user, all of the toll lane trips have been taken: the corresponding cumulative trip count percentage for the 50th percentile user is 100%. The GTRI-matched users in the second chart differ more at the high end: here, the top 10% of users took 70.69% of the Express Lane trips, while the 50th percentile of users took 98.49% of the toll lane trips. Finally, at the demographic-matched level, the top 10% of users accounted for 58.20% of all Express Lane trips in 2013, while the top 50% accounted for 97.37%.

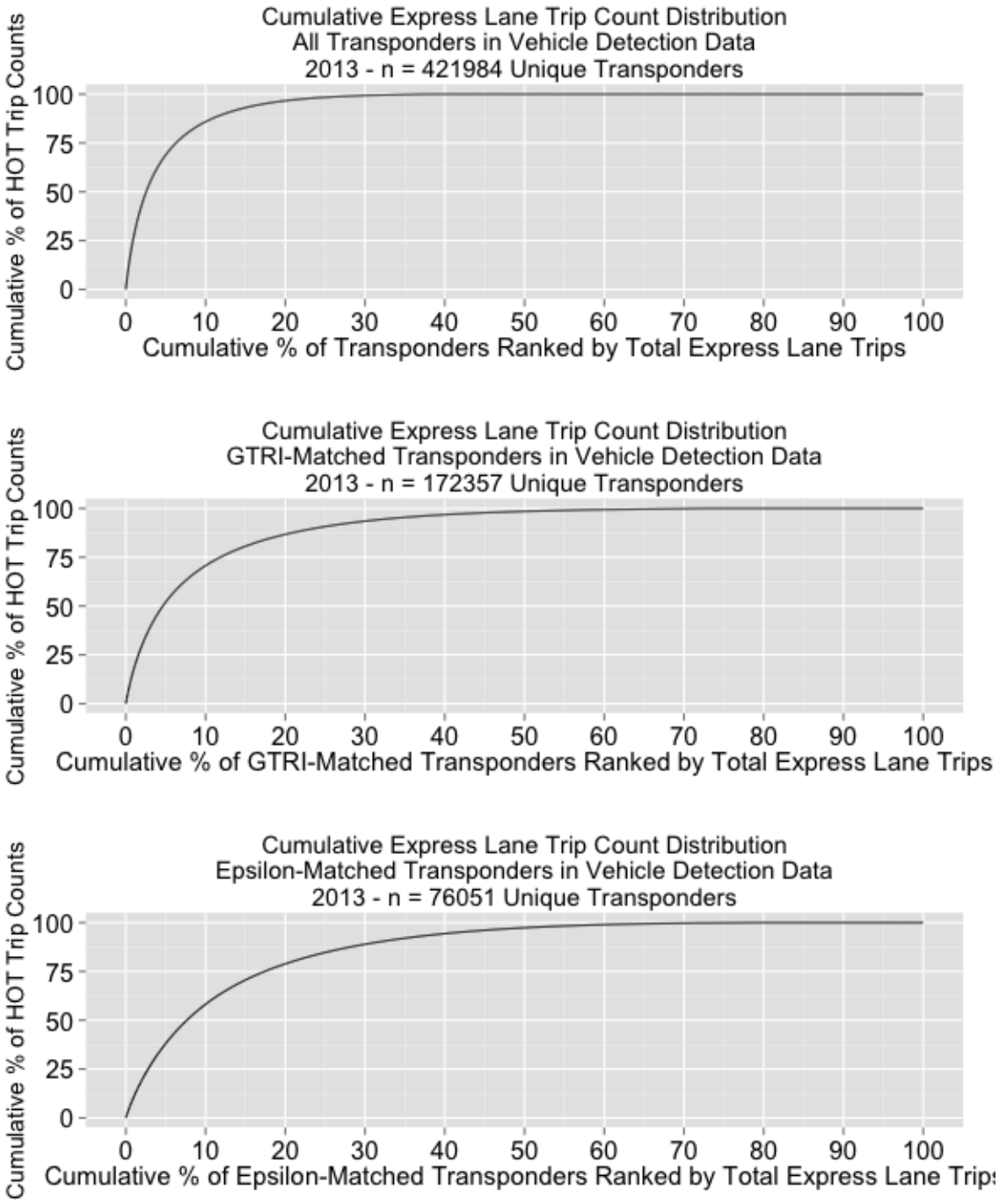


Figure 98: HOT Trip Count Distribution by Pairing Level

Iterating through the data pairing processes reduces the total proportion of corridor trips taken by the top 10% of users by an average of 12.78% at each of the two stages. At the median, each pairing stage reduces the percentage of trips taken by 1.87%.

For the toll lane trips, these averages are slightly higher: the top 10% of users see an average of 13.84% reduction in their share of trips taken at each stage of the process. The top 50% of users see their total trip percentage decrease by 1.32% at each stage on average. These rates are understandable, as by definition the most frequent users take the most trips; removing more frequent users will therefore have more of an impact on trip counts than removing less frequent users.

While the cumulative trip count distribution charts show some of the impact of the pairing process, the issue that arises with these figures lies in the uncertainty in the loss of trip data. That is, the plots do not indicate whether the data loss was random or whether it was concentrated among certain users. The flattening plots show that frequent users drop out of the sample during the pairing process: while the top 10% of all corridor users took nearly 70% of the corridor trips in 2013, the top 10% of demographic-matched users took less than 45% of the demographic-matched trips. Less apparent is the impact of the pairing process on less frequent users, or the distribution of the impact on frequent users relative to those less frequent users. The next section seeks to address this shortcoming.

Pairing Dropouts by Rank

To address the limitations of the cumulative distribution plots discussed above, trip data loss was assessed as a function of user rank. As in the previous section, rank here is defined by a user's position within the list of transponders ordered by trip count. The purpose of this investigation is to examine whether the data loss incurred during the pairing process is randomly distributed among corridor users or whether it is concentrated

within a specific portion of the transponder population. This section examines the paired dropouts by their ranks before and after the pairing process.

Y-Y Plots of Changes in Rank

The first method used to investigate this question compared the transponder ranks before and after the two steps of the pairing process. After ordering the transponders by the number of trips taken per transponder, each transponder was assigned a percentile rating based on its rank. The most frequent trip takers, for example, were found in the first percentile. The list of transponders was then narrowed to those which could be paired with GTRI registration database addresses, and also those whose addresses placed them in the I-85 commutershed. The author assigned a new set of ranks to this new list, so that each paired transponder had a pre-pairing rank and a post-pairing rank.

Figure 99 below is a Y-Y plot illustrating the percentile ratings, based on the number of total corridor trips for each transponder, of the sample of all transponders detected in the constructed trips versus the sample of address-matched commutershed transponders in the constructed trips. A transponder's position on the x-axis indicates its percentile rank in the original unpaired data set, while its position on the y-axis shows its percentile rank in the GTRI-matched commutershed dataset. The trip counts and transponder lists were taken from the duration of calendar year 2013. The plot shows that a user at the 25th percentile of trip frequency within the full dataset is ranked at 45.10% in the GTRI-matched commutershed dataset. A user at the 75th percentile in the full set has a corresponding ranking in the matched dataset of 93.05%. Additional percentile rankings can be found in Table 35.

The shape of the curve in Figure 99 illustrates the nature of the data loss. If the losses were randomly distributed among the users, the resulting curve would follow the y = x line in the figure. As greater percentages of data go missing, the y-y curve departs from the straight line. The 25th percentile figure mentioned above, which yields a 45.10% rank in the GTRI-matched set, indicates the loss of less frequent users. Similarly, the 50th percentile user appears at the 75th percentile in the GTRI-matched data. The users ranked below the 50th percentile have suffered more data loss than those above the 50th percentile, and so the user's relative position in the matched rankings decreases. As more data are retained, the curve arcs back towards the straight line.

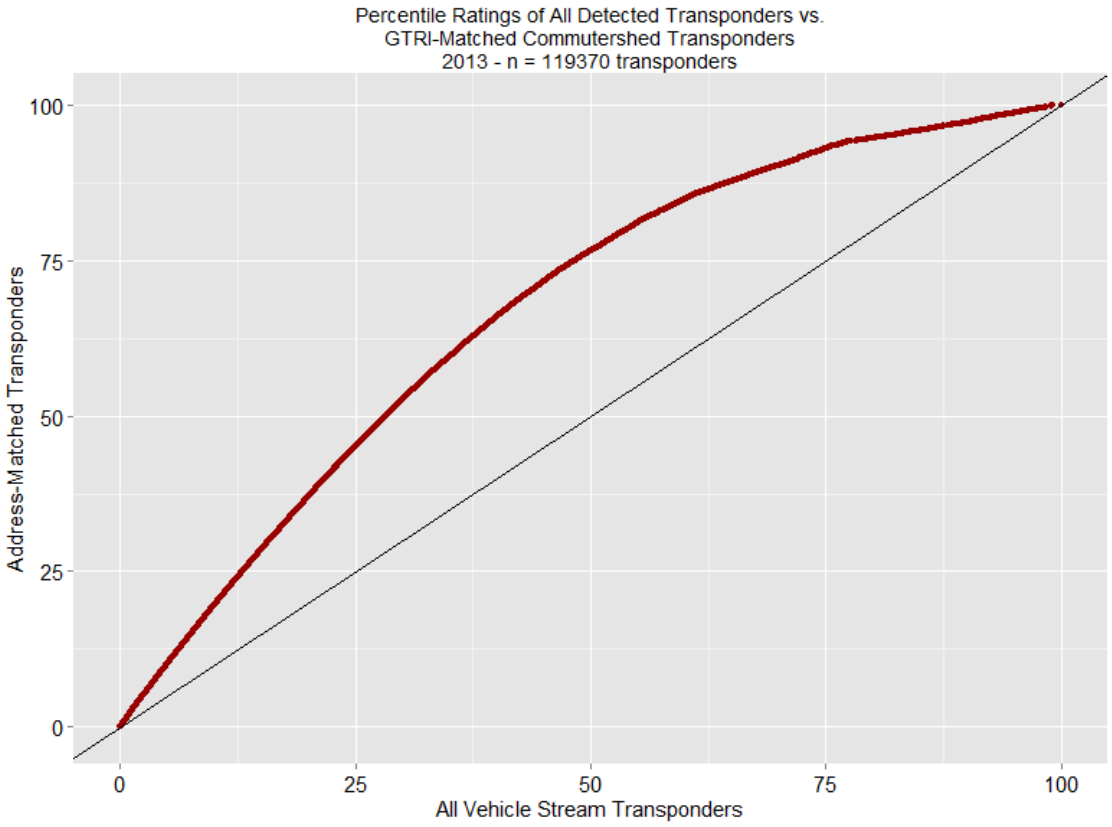


Figure 99: Percentile Ranks - All Detected vs. GTRI Commutershed Transponders

Figure 100 shows a similar Y-Y plot comparing the relative ranks of all of the transponders detected in the constructed trip dataset versus the subset of commutershed transponders for which an Epsilon demographic match could be made. The y-axis in this plot shows the transponder's percentile ranking within the Epsilon-paired commutershed dataset. Here the 10% rank within all transponders corresponds to a 34.61% rank among the demographic-matched commutershed transponders. At the 50% level within all transponders, the matching Epsilon-matched rank is 89.03%. This plot shows more significant differences between the original and paired ranks relative to the GTRI-matched chart; the differences in rank are greater at virtually every point across the spectrum.

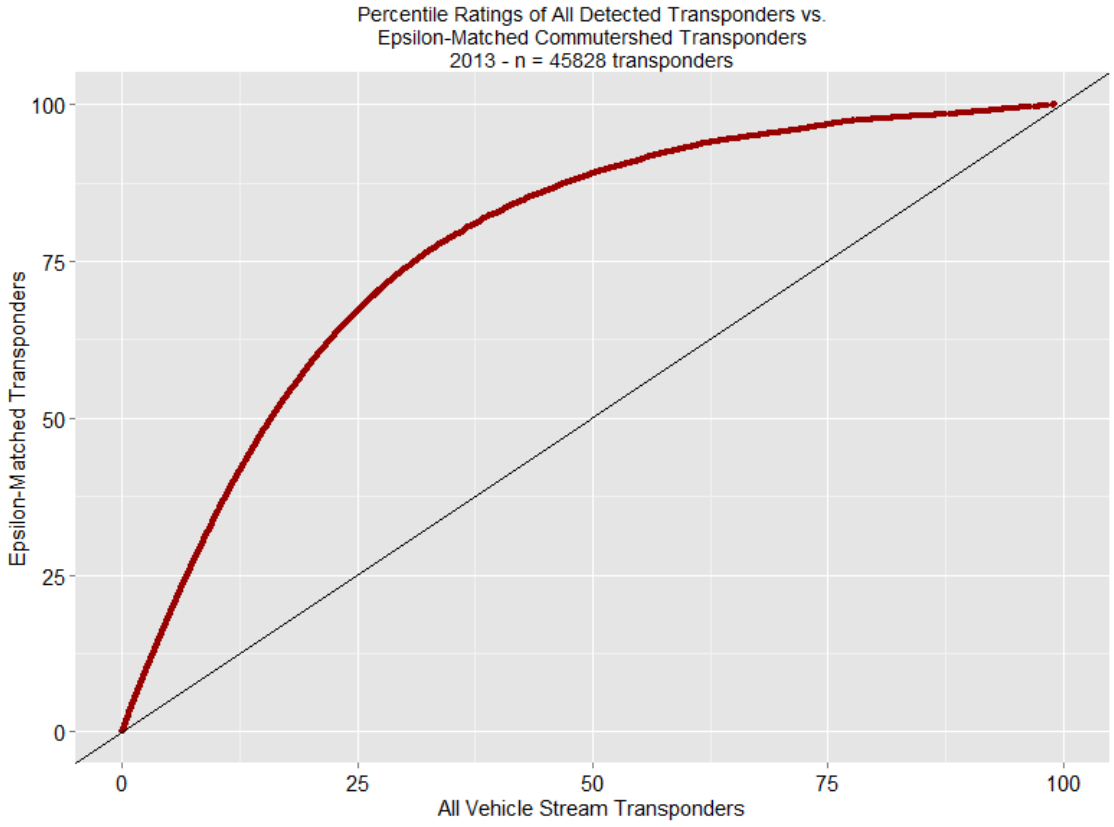


Figure 100: Percentile Ranks - All Detected vs. Demographic-Matched Commutershed Transponders

Table 35 provides an overview of the percentile ranks in the original, GTRI-matched, and Epsilon-matched data sets. At every percentile level in the table, the corresponding GTRI and Epsilon matched ranks are significantly lower (that is, the percentile values are higher in the paired sets). This indicates that there were fewer users below (with a higher percentile rank) a given transponder in the paired data. For example, the 10th percentile transponder became the 19.62% transponder by virtue of users in the 11-99th percentile group dropping out of the GTRI-paired data set.

The largest discrepancy between the original rankings and the GTRI-matched rankings shown in Table 35 occurs at the 50% level; the difference between the percentile ratings at that point is 26.58%. Expanding the search to include percentile values outside of the table, the largest discrepancy across the whole spectrum occurs at the 46.78% rank in the original data set. The corresponding GTRI-matched transponder rating at that point is 73.58%. Between the original data set and the Epsilon-paired sample, the largest gap in the rankings presented in the table is at the 25% level: there the difference is over 40%. After this point, the rankings begin to converge again as sample retention improves. The largest gap overall occurs at the 32.10% position in the unpaired sample; the percentile rating for that transponder in the demographic-matched set is 76.13%. In both the GTRI- and Epsilon-matched rankings, the smallest difference is at the 90th percentile.

The main takeaway from this table is the indication that more frequent users are more represented in the paired data sets. At each rank level examined in the original data set, more of the less-frequent users are dropped relative to more-frequent users. This effect is more pronounced among the demographic-matched transponders than it is

among the GTRI-matched transponders. As in the previous section, these charts do not tell the whole story concerning which users are dropped. The next section will examine the number of dropouts at each percentile rank in the various data sets.

Table 35: Percentile Ranking by Pairing Step

Full Transponder Set Ranking	GTRI-Matched Commutershed Transponders	Demographic-Matched Commutershed Transponders
10%	19.62%	34.61%
25%	45.10%	67.06%
50%	76.58%	89.03%
75%	93.05%	96.76%
90%	97.30%	98.82%

Dropout Counts by Rank

While the previous section presented the relative ranks by total trip count of transponders before and after the demographic pairing process, it did not delve into the details behind the changes in those ranks. This section seeks to expand upon that analysis by examining the numbers of transponders that are lost in the pairing process as a function of the frequency of their trip-taking. The figures below present two perspectives on this issue.

Figure 101 shows the number of dropouts that occur in the marketing data matching process at each transponder percentile rating. Here again, the transponders were ordered by the number of trips they took in 2013. Each transponder was then assigned a rank and corresponding percentage value based on the total number of transponders. In a 100-transponder sample, for example, the transponder with the most trips would be assigned to the 0-1% bin, represented here by a percentile rating of 0%. In this sample, each percentile bin contained 4,219 or 4,220 transponders. These bins are represented on the x-axis of the chart below. The y-axis displays the number of transponders from the original data set that were dropped during the demographic data

matching process. For example, the first bin (at 0%) lost 1,320 out of 4,219 transponders, or 31.3%, after the data were paired to the GTRI and then the Epsilon data.

What is immediately apparent in the figure is the increasing dropout rate among higher percentile ranks. That is, the number of dropouts per percentile bin increases as the number of trips represented by each bin decreases. The last quartile of bins consist almost entirely of dropouts; few if any transponders from those groups are present in the demographic-matched data. Note that these losses may be due to a number of reasons: the households may be located in an area for which no marketing data were purchased (this is likely the case for those users with Georgia State Route 400 toll tags that do not live in the I-85 commutershed), there may have been an error in the addresses used for data set matching, or they may be less-frequent users which were not originally targeted in the marketing data purchase.

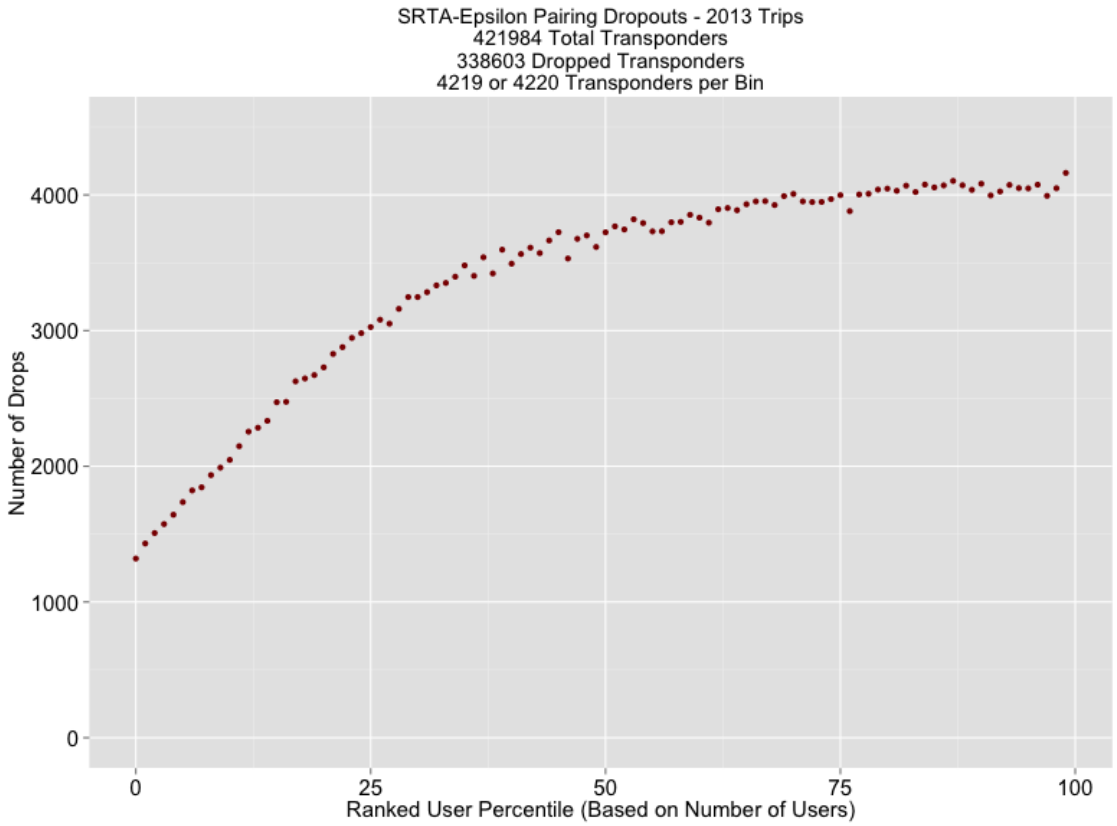


Figure 101: Paired Dropouts by User Percentile

Figure 102 differs in that the percentile rating for each user is calculated by the total number of trips, not the total number of users. This has the effect of changing the number of transponders represented in each percentile bin. The 0-1% bin, for example, represents 180 transponders that collectively took 1% of the total corridor trips in 2013. The 99% bin, on the other hand, includes 116,883 transponders, each of which took an average of only 1.2 trips in all of calendar year 2013. While the scale of the chart flattens the losses of the first three transponder quartiles, the losses in the remaining 25% are striking. The results are consistent with the previous figure: transponders with fewer trips are less likely to appear in the demographic-matched data set.

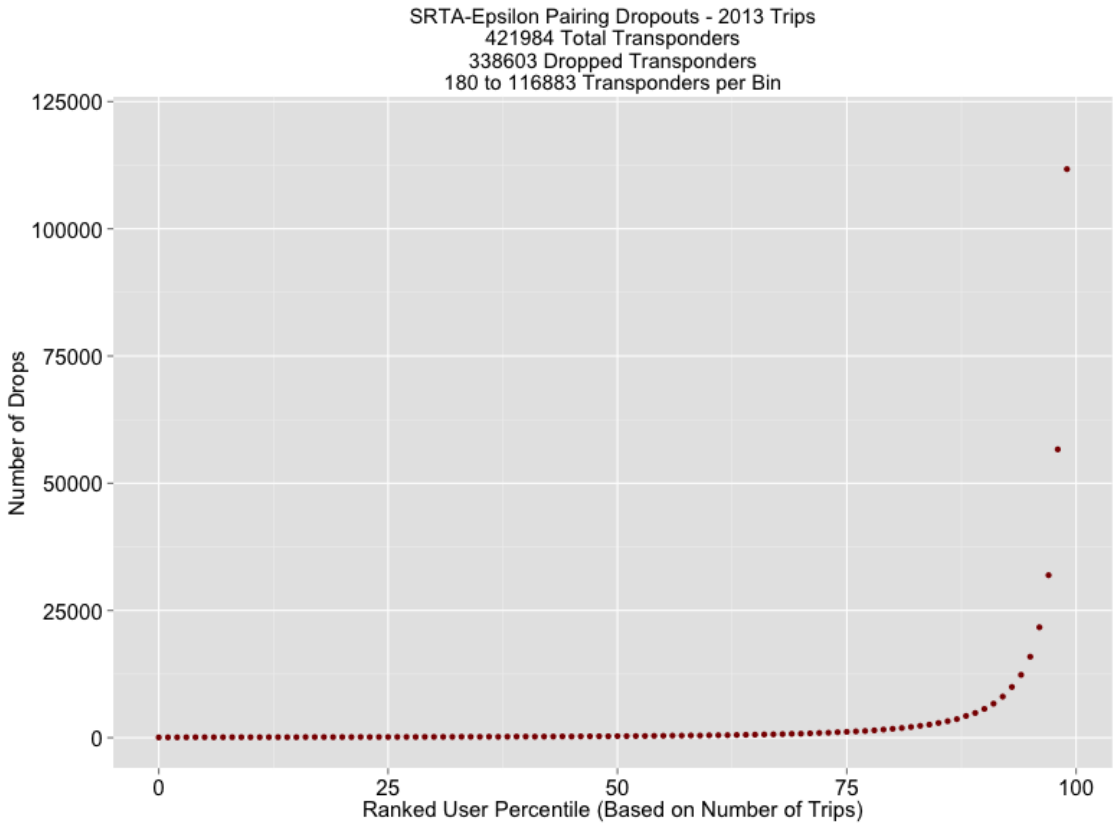


Figure 102: Paired Dropouts by Trip Percentile

The pattern that emerges from both of these figures is the bias in the pairing process towards frequent users of the I-85 corridor. The households for which the Epsilon demographic data was purchased, and to a lesser extent the households which can be paired to GTRI registration database records, are those which more frequently use the I-85 corridor and the Express Lanes. This is to be expected, as the Epsilon marketing data purchase was deliberately targeted towards vehicles that were more frequently observed on I-85.

Census Data Comparison

Because demographic data from the purchased marketing data source were not available at all levels of the data pairing process, this research used Census data from the

American Community Survey 5-Year summary file to examine the demographics of the households at various stages of the pairing process. The 5-year summary file was selected for its geographic specificity; it is the only summary file to present data at the block group level (U.S Census Bureau, 2013). This chapter uses the 2009-2013 ACS Survey as it is the most recent version available at the time of writing.

The first step in this process involved geocoding the results from the Georgia registration database matching process. Of all 983,860 plates sent to GTRI for registration database matching, 518,169 (52.7%) were returned. Among these returned records were 366,298 unique households. An address locator for the Atlanta region was constructed in ArcGIS using Census TIGER street data (United States Census Bureau, 2014). These county-level street data, provide by the Census bureau, were combined by the Atlanta Regional Commission (ARC) and included in their Atlanta Regional Information System (ARIS) data set (volume 1c, 2011). After constructing the address locator, the author geocoded the 366,298 addresses matched to the registration database. 293,883 of those addresses were successfully geocoded. These geocoded results were then spatially joined to the Census block group in which they reside. Figure 103 illustrates the geocoded address matches and the Census block groups which contain them.

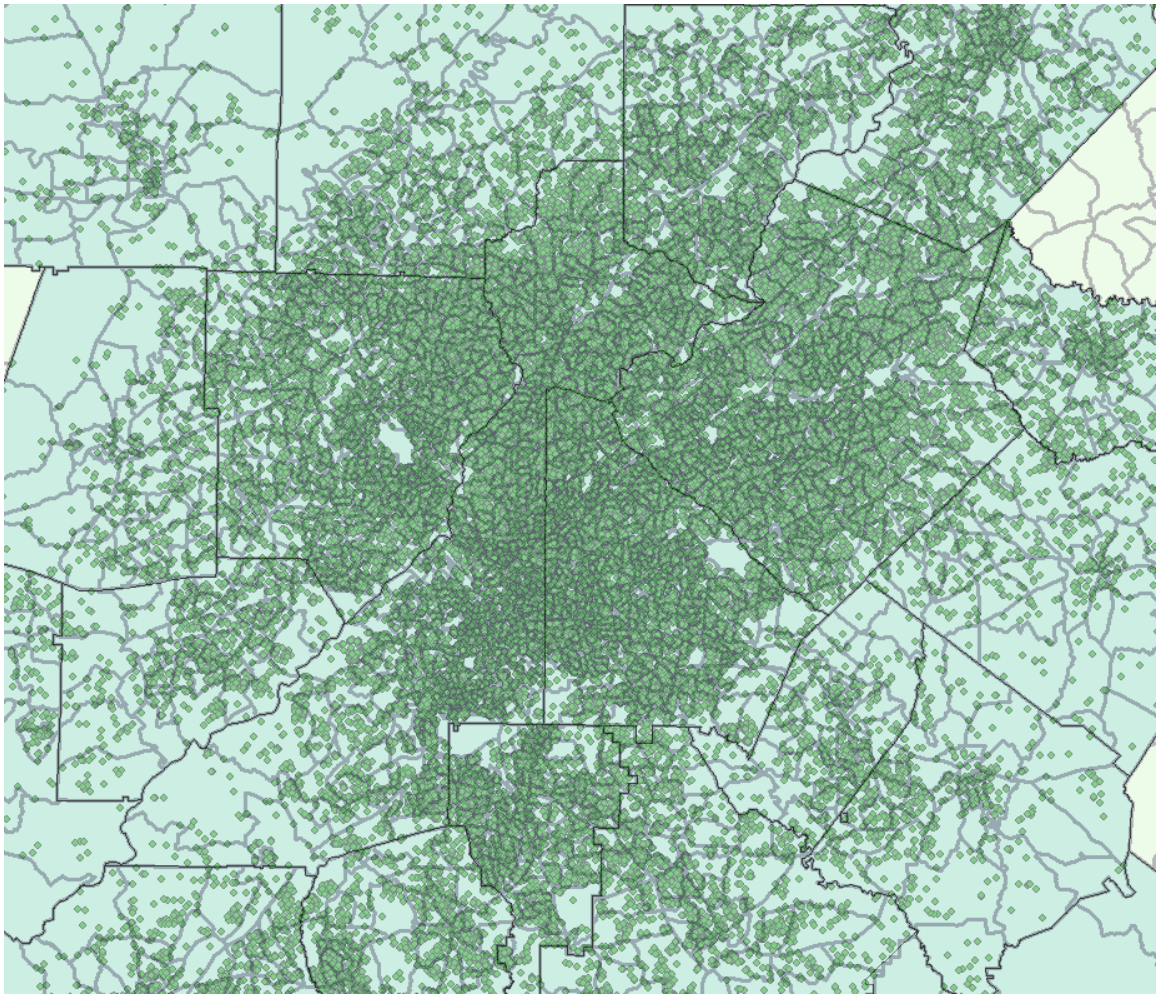


Figure 103: Geocoded Address Matches with County and Block Group Boundaries

Once the geocoded addresses were joined to the Census block groups, the author examined the American Community Survey 5-year data for those block groups to illustrate the income distributions of matched households. Figure 104 below illustrates the distribution of census block group median income values for the 293,883 geocoded households in the registration-database matched data set. These geocoded households represent 80.2% of the 366,298 households matched to the registration database. This chart is presented at the top of Figure 104; the median value (of the block-group level median values) is just over $70,000 per year.

The second plot in Figure 104 presents the distribution of ACS block group median incomes for the geocoded households from the Epsilon marketing purchase dataset. This dataset was generated separately from the previous registration-matched dataset, which used the SRTA trip records and account data as the source of its license plates. The source of the license plates for the Epsilon demographic purchase was the two-year HOV-to-HOT conversion analysis that Georgia Tech conducted from 2010 to 2012 (Guensler, et al., 2013). This project involved the collection of 1.5 million license plates of I-85 corridor users. Though two sets of license plates were collected using different methods (video observation versus automated reporting), there is significant overlap among them.

Of the 349,134 households in the purchased marketing dataset, 289,557 (82.9%) were successfully geocoded. The second plot in Figure 104 shows a different distribution shape for the geocoded Epsilon households compared to the geocoded GTRI-matched households. Here the median household income is almost $8,000 lower, and the whole distribution is shifted to the left (towards lower incomes). The final chart in Figure 104 illustrates the households from the Epsilon marketing data that were successfully paired with the SRTA transponder data. This pairing process is described in detail in the Connecting SRTA Data to Epsilon Data chapter. This figure used December 31, 2013 as the date on which the SRTA and Epsilon data were paired. A total of 40,426 households were successfully matched. The resulting distribution of ACS median incomes is higher than both of the previous sets, with a median of over $76,000. The distribution is also less heavily tilted towards the lower end and exhibits more of a rightward-shift, towards higher incomes, than either of the two previous charts. These results suggest a noticeable

bias towards higher incomes in the SRTA-Epsilon paired data. An overview of these three distributions is provided below in Table 36.

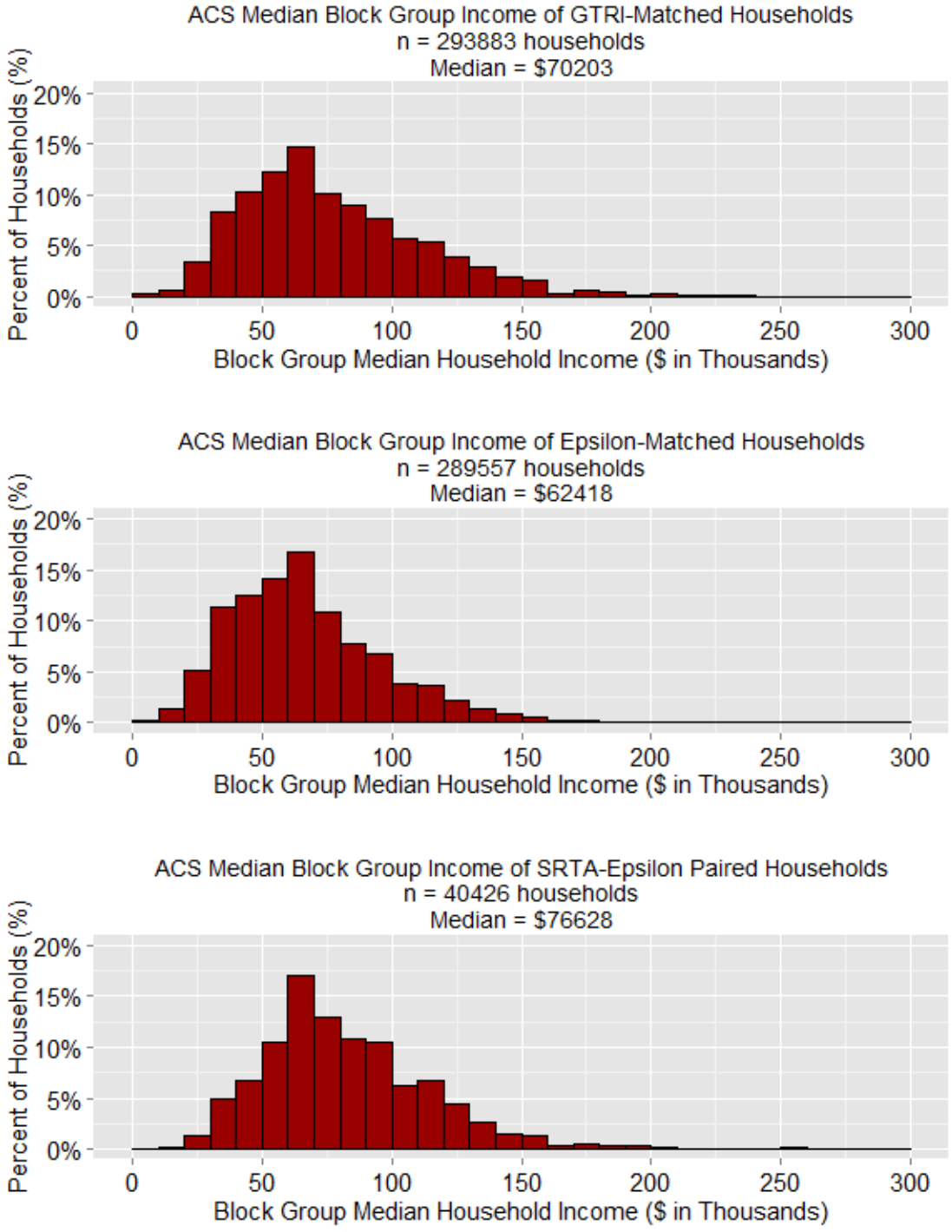


Figure 104: Median Census Household Income Distributions

Figure 105 presents distributions of the ACS block group household age data for the same three sets of households. The first plot illustrates Census-provided household age data for registration database-matched households. Here the median value of the ACS data is nearly 37 years old for the 293,883 households in the sample. Again, the values reported for each household are the ACS estimates of the median household age of the block group in which the household is located.

The second and third charts in Figure 105 present the demographic-matched households and the SRTA-Epsilon paired households, respectively. Within the demographic-matched sample, the median head of household age drops slightly, by less than one year. The SRTA-Epsilon paired households have the same median age as the GTRI-matched households, though the shape of the distribution differs. The paired sample, which is ultimately used in the analyses in this dissertation, has a more-concentrated peak around the median, with fewer households on the shoulders of the distribution. These charts indicate that the paired dataset, while exhibiting similar central household age measures, include marginally fewer households at the shoulders of the distribution.

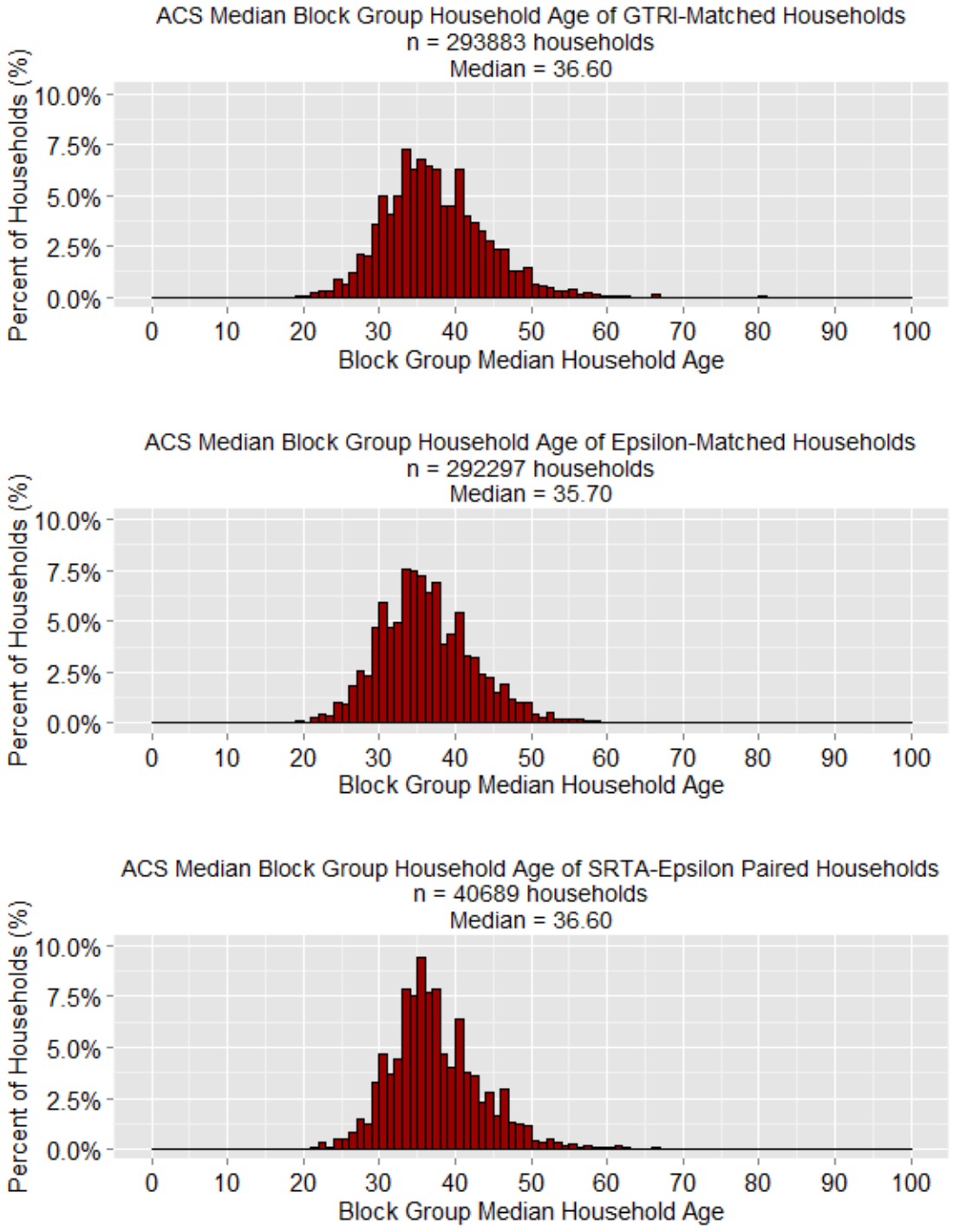


Figure 105: Median Census Household Age Distributions

Figure 106 presents the final set of distributions for the GTRI-matched, demographic-matched, and SRTA-Epsilon paired households: that of average household size of family households. For each block group, the ACS reports estimates of the numbers of households with two individuals, three individuals, and so on, up to seven individuals. The ACS also splits these estimates into 'family households' and 'non-family households.' The distributions presented in Figure 106 include family household data. The average household size for each block group was computed by counting the total number of persons reported by each family household category and dividing that value by the number of family households. The inclusion of only 'family households' explains the minimum household size value of 2. Here the average household size distributions for each sample are similar, with median values that differ by 0.07 at most. While the GTRI-matched sample has a less-pronounced peak than the rest, the shapes of the distributions are otherwise alike. The activity-demographic paired sample has the highest median household age, but the magnitude of the difference and the similarity in the shapes of the distributions suggest that the discrepancy is not significant.

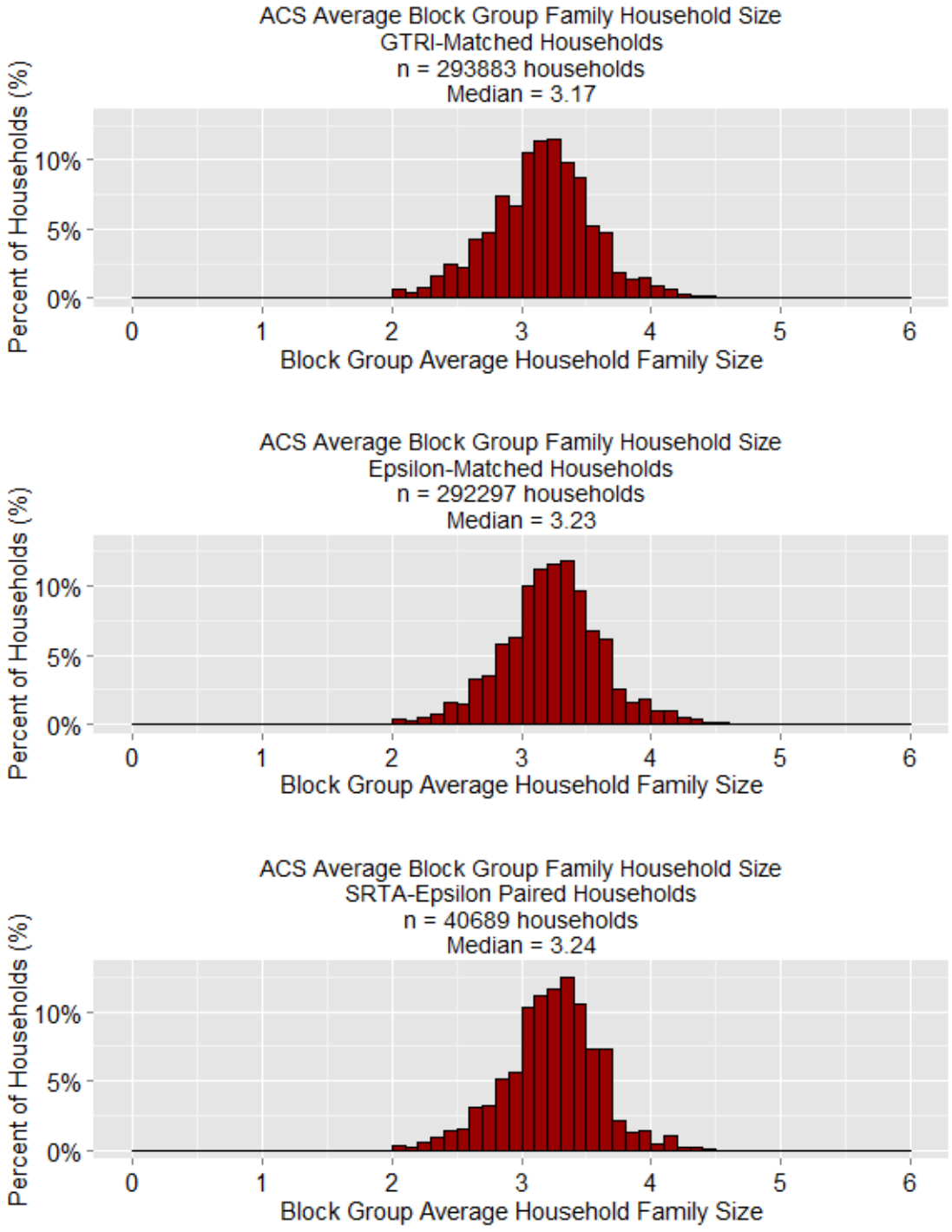


Figure 106: Average Census Family Size Distributions

Table 36 provides an overview of the census data distributions for all three data sets. The most striking differences appear in the household income comparison. The average household income in the paired sample is over $15,000 higher than that of the full set of Epsilon households (which is not surprising as it includes non-freeway-users), but is also $5,000 higher than that of the GTRI-matched households, which might indicate a bias if the dropout rate is higher for low income households. On the other hand, it may simply reflect that lower ACS income areas have a lower fraction of household users of the facility and that the census block average income does not reflect the average income of users from that census block. The demographic-matched households have the highest level of kurtosis and the smallest inter-quartile range between the 25^{th} and 75^{th} percentiles. As suggested by the distribution figures above, the SRTA-Epsilon paired households have substantially higher annual incomes than the larger pools from which they are drawn.

The household size distributional measures reflect the similarities apparent in the figure. The average and median household size values are similar across all three data sets, as are the inter-quartile range measures. The skewness results indicate that all three distributions are close to symmetrical, while the kurtosis values indicate that all three are more peaked than the normal distribution.

The household age distributions are also more similar than different. Here the SRTA-Epsilon paired distribution is slightly less symmetrical, and slightly more peaked, than the other two samples. Other measures also differ only marginally: the average household age in the SRTA-Epsilon paired data is 1.13 higher than that of the demographic-matched data, and the inter-quartile range is smaller than that of the GTRI-

matched data by 1.1. In all cases, Mann-Whitney tests rejected the null hypothesis of distributional equality at the 99% confidence level; but, this was to be expected given the large number of observation counts in each of the three data sets.

Table 36: Overview of Census Data Distributions

	GTRI-Matched Households	Demographic-Matched Households	Paired Epsilon Demographic Sample
Number of Households	293,883	289,557	40,426
Household Income			
Mean	$77,078	$67,012	$82,201
Median	$70,203	$62,418	$76,628
25th Percentile	$51,458	$45,380	$60,625
75th Percentile	$97,763	$82,317	$99,891
Skewness	0.932	1.067	1.037
Kurtosis	4.171	5.097	5.067
Mann-Whitney Test Result: vs. GTRI	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Mann-Whitney Test Result: vs. Epsilon	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$
Mann-Whitney Test Result: vs. Paired	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A
Household Size			
Mean	3.162	3.242	3.238
Median	3.174	3.232	3.240
25th Percentile	2.907	3.020	3.051
75th Percentile	3.408	3.478	3.478
Skewness	0.0343	0.0870	-0.187
Kurtosis	3.564	3.807	3.712
Mann-Whitney Test Result: vs. GTRI	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Mann-Whitney Test Result: vs. Epsilon	$p<2.2x10^{-16}$	N/A	p=0.0065
Mann-Whitney Test Result: vs. Paired	$p<2.2x10^{-16}$	p=0.0065	N/A
Household Age			
Mean	37.29	36.23	37.36
Median	36.60	35.70	36.60
25th Percentile	32.80	31.80	33.40
75th Percentile	41.20	40.10	40.70
Skewness	0.628	0.605	0.758
Kurtosis	4.277	4.172	4.672
Mann-Whitney Test Result: vs. GTRI	N/A	$p<2.2x10^{-16}$	p=0.0046
Mann-Whitney Test Result: vs. Epsilon	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$
Mann-Whitney Test Result: vs. Paired	p=0.0046	$p<2.2x10^{-16}$	N/A

Account Stream Join Issues

As discussed in previous chapters, the registered account data that Georgia Tech receives from ETCC contains a many-to-many relationship between the transponders and vehicles. That is, there is no linking element within an account's record that specifies which transponder is associated with which vehicle license plate. As a result, accounts with multiple vehicles and multiple transponders do not identify which vehicle is using which transponder. This is a complication in the analytical process, as the method of pairing SRTA records with demographic data involves household addresses matching. So if an account has multiple registered vehicles, which are registered at different addresses, the results would link to more than one demographic data set.

Initial analyses reported in Sheikh (2016) addressed this issue by examining only those accounts with a single transponder and a single registered vehicle which allowed for a direct link between the transponder, the license plate number, the registration address, and finally the demographic data. This method introduced bias into the results by including only accounts and households with a single vehicle. An examination of the Account stream data, shown below in Table 37, shows the number of active accounts with zero, one, and two or more matching Epsilon records (by column) on January 1, 2014. These are accounts that are not in 'Closed' status (the remaining status values include 'Active,' 'Proposed,' and 'Pending to Close'). The bottom row shows the total number of transponders associated with those accounts. In previous analyses, the majority of accounts, 87.2%, were not matched with a corresponding Epsilon ID and so cannot be paired with demographic data. Of the 278,170 accounts, 230,503 (82.9%) were not paired with any Epsilon records. 45,955 (16.5%) were paired with one marketing

record, and 0.62% were paired with multiple marketing records. These accounts, for which multiple marketing records were found, cannot be included in the analyses because they cannot be tied to a specific household. Restricting the Account data to accounts that were not closed reduced the rate of unpaired records by 4.3% and increased the rate of accounts paired to one Epsilon record by 4%.

Table 37: Counts of Active Accounts with Matching Demographic Data IDs

Number of Active Transponders	Number of Accounts Matched to 0,1, or 2+ Demographic IDs (percent of row total)		
	0 IDs	1 ID	2+ IDs
0	22,035 (100%)	0 (0%)	0 (0%)
1	126,896 (85.3%)	21,758 (14.6%)	102 (0.1%)
2	58,491 (78.7%)	15,290 (20.6%)	516 (0.7%)
3	15,570 (72.5%)	5,539 (2.58%)	359 (1.7%)
4	4,805 (69.3%)	1,924 (27.7%)	205 (3.0%)
5	1,551 (66.2%)	682 (29.1%)	109 (4.7%)
6+	1,155 (49.4%)	762 (32.6%)	421 (18.0%)
Total Active Transponder Count	329,246 (75.3%)	88,602 (20.3%)	19,517 (4.5%)

Data Pairing and Join Loss

Every step in the dataset construction process entails some degree of loss. This loss occurs when a join cannot be made successfully for a trip due to a lack of data. At the demographic data join stage, for example, a trip is excluded if the transponder could not be paired with a unique demographic record. For other stages, such as the toll rate or travel time joins, trips are excluded when no corresponding records can be found in the toll rate or travel time databases that match the date, time, and location of the trip in question. The purpose of this section is to list the steps in the process and quantify the data loss that occurs at each of those steps. Table 38 presents an overview of the joining process for the month of January, 2013, which includes more than one million constructed trips made by 120,822 unique transponders.

The first step is pairing the constructed trips with demographic data, which eliminates a significant portion of the available trip data (46.24% of the constructed trips from January 2013 cannot be matched to demographic data). This step also removes an even larger portion of the transponder population, 64.88%. In terms of trip characteristics, the changes are primarily in the category of lane type.

After the vehicle activity data are joined with demographic records, the percentage of trips by lane use also change, as shown in Table 38. General Purpose lane-only trips decrease from 65.8% to 59.9%, while the Express Lane-only trips increase from 12.4% to 14.3% after matching, and the mixed-lane trip percentage changes from 21.8% to 25.8%. Average trip speed decreases slightly after matching, and while the average trip length is greater, than measure is problematic because it is correlated with lane use (as discussed previously in this dissertation).

The next three steps in the join process (trip stream join, the toll rate stream join, and the average travel time join), have a much smaller impact on the numbers of trips and transponders in the data sets. Of these three stages, the travel time join has the largest effect on the sample. This join reduces the number of trips in the data set by almost 100,000, and increases the general purpose-exclusive trip rate by over 4%. Average trip speed after the travel time join is only marginally different than the previous three stages.

After the demographic data join stage, the stage with the largest impact on the data set is the transponder detection count join. This join reduces the trip count by over 50% relative to the previous travel time join stage; the resulting count is 21.2% of the original sample. Similarly, number of transponders in the sample decreases by roughly 30% relative to the previous stage, and the remaining 28,740 transponders account for

23.8% of the initial data set. Compared to the original, unjoined data set, the lane type use rates differ greatly. The number of trips that occur solely in the general purpose lanes decrease from 65.8% to 53.3% compared to the original constructed trips; most of that difference is balanced with an increase in the rate of mixed trips.

The loss of data during the data set construction process is particularly significant at the stage in which vehicle activity is joined to marketing data (loss of 46.2%) and then the transponder detection count join stage (loss of an additional 32.5% of the original data). The final data set consists of just 21.2% of the original trip count, and 23.7% of the original transponder count. The question that naturally arises from this is whether sample biases result from exclusion of data from the individual join processes. While the loss of the marketing data is largely the result of households from outside the commutershed (38.7% of the active transponder population consists of transponders from the now-defunct Georgia 400 toll, a facility with a different catchment area), and are largely associated with infrequent users, the transponder detection count join losses are less intuitive. It may be more valuable or worthwhile to forgo this step in favor of preserving more of the constructed trips. As later chapters will demonstrate, this question is complicated by the fact that the transponder count join is the basis for some of the most valuable model components. Similarly the Epsilon demographic data join, which removes the largest number of trips and transponders, is required for the demographic analyses that motivate this entire dissertation. As a result, the strategy employed here is to describe the impacts of this data loss in this and other chapters of the dissertation so that the analytical results can be interpreted with full knowledge of their limitations.

Table 38: Data Loss by Join Step - January 2013

	Constructed Trips	After Demographic Join	After Trip Stream Join	After Toll Stream Join	After Travel Time Join	After Transponder Count Join	After Account Join
# Trips	1,076,511	578,724	543,079	531,630	464,487	228,463	228,060
# Transponders	120,822	42,438	41,978	41,897	40,957	28,740	28,673
Avg. Length (mi)	8.21	8.59	8.46	8.47	8.28	8.88	8.89
Avg. Speed (mph)	63.8	62.5	62.7	62.9	62.6	56.3	56.3
% HOT Trips	12.4	14.3	13.1	12.9	9.7	14.6	14.6
% GP Trips	65.8	59.9	63.8	64.1	68.5	53.3	53.3
% Mixed Trips	21.8	25.8	23.1	23.0	21.8	32.1	32.1

Demographic Characteristics of Paired Data

This chapter has focused on the impacts of the Express Lane use and demographic data pairing process, specifically as it affects the transponder population and the trip characteristics of those transponders. Another very important category of those impacts is the demographics themselves; that is, how the demographic characteristics of the sample change throughout the pairing and data set construction process. This issue is discussed in Chapter 5. Whereas this chapter uses Census data for the paired households, Chapter 5 (Connecting SRTA Data to Epsilon Data) examined the Epsilon data set as a whole. The investigation of the paired Epsilon households revealed a sample that had higher average household incomes, older heads of household, larger household sizes, and slightly higher education levels than the full marketing data purchase. Chapter 6 (Data Quality and Treatment) outlines issues with the demographic data, including the mishandling of multi-family dwelling units (which was corrected). Chapter 6 (Connecting SRTA Data to Epsilon Data) chapter also compared the users in the Epsilon-paired sample with Census Bureau data for the City of Atlanta reported in the five-year American Community Survey. Those differences included higher median household incomes in the Epsilon sample, along with fewer single-occupancy households, more undergraduate degrees, and far more home owners. The ACS comparison is less than direct, however, as the measures were taken at different levels (households for the Epsilon data, individuals for much of the Census data) and different geographies (the I-85 commutershed for the Epsilon data, the City of Atlanta for the Census data).

Chapter Summary

The purpose of this chapter was to examine the different ways that manipulation of the lane use and demographic data, primarily through the process that paired the two disparate data sets, had the potential introduce bias into the analytical results. The mechanisms that created the possibility for bias include matching the SRTA corridor use data with the vehicle registration database, matching those results with the Epsilon demographic data purchase, and constructing the complete data set including corridor operational characteristics. The impacts of these data processing stages were seen in the subset of Peach Pass transponders and Epsilon households that made it through the entire process. The resulting sample differed from the complete set of SRTA data by primarily including those vehicles that frequently used the corridor; the bottom quartile of users ranked by trip frequency (infrequent users) were virtually excluded from the final paired sample.

The effects of the data processing required for the analyses in this dissertation on the demographics of the sample were examined in different ways. Chapter 5 compared the paired demographic data with the full data purchase. That chapter also compared the paired households with City of Atlanta dwellers using Census ACS data. Chapter 7 compared the ACS-provided demographic characteristics of the GRTI-matched households, to the households for which demographic data were procured, and with the households for which the SRTA-Epsilon pairing was successful. That investigation found a substantial bias in the SRTA-Epsilon paired sample towards higher income households, while the other demographic characteristics examined were largely similar. However, it is impossible to know whether the difference constitutes a sample bias or a

revealed difference between facility users and non-users at varying levels of income aggregation (household vs. census level).

This chapter also examined the data loss that occurred in joining the SRTA constructed trips with the Epsilon demographic data and with the other streams that were provided by SRTA or derived from their data. The join process results in the exclusion of a significant portion of the constructed trip population; the trips that remain at the end of the process differ primarily in the higher rates of toll lane use, lower average speeds, and fewer households represented. The structure of the Account data stream is another potential source of bias: left unaddressed, the many-to-many relationships in the data stream restrict analysis only to those accounts with a single transponder and vehicle. Expanding the analysis to accounts with a single household address illustrates the potential scope of this bias: of the 88,602 active transponders associated with a single Epsilon record, only 21,758 (24.56%) come from single-transponder/single-vehicle accounts. These issues must all be weighed and considered when conducting and interpreting the analyses that are performed later in this dissertation.

CHAPTER 8

INITIAL HOT USE CHOICE ANALYSIS

This chapter begins the investigation into the behavior and decision making of users and non-users of the I-85 Express Lanes, within and across various demographic groups. The chapter uses the unique combination of Express Lane data and household demographic data to examine decision-making at the trip level for users from different income groups and demographic clusters. Additionally, this chapter provides a summary of Express Lane use characteristics by these different groups. The results may help inform future toll lane studies and investigations of equity impacts.

The next section describes the sources of the data used in the study and provides an overview of the dataset. The methodology section explains how the data were processed and the modeling techniques that were applied. Next, the results section discusses the modeling outputs. Finally, the chapter addresses the limitations of the data employed in the study and describes the next steps in this research. A version of this chapter was submitted to the Transportation Research Board for the 2015 Annual Meeting; the paper was selected for presentation and for anticipated publication in 2016 (Sheikh, et al., 2015).

Data

The data supporting this initial study come from the two sources described previously: Express Lane use data collected by SRTA, and household socioeconomic data procured from Epsilon, a marketing firm. This chapter uses vehicle detection and toll rate data for calendar year 2013. Within that year were over 157 million transponder detections and

more than 100,000 toll updates for the year (five-minute intervals). Vehicle detections originate from the various HOT and GP lane RFID detectors. Each detection record provides the unique transponder identification number, the detection time, and the lane type and gantry number at which the detection occurred. The toll amount data stream was also used to identify the posted toll rate for each HOT entry and exit combination throughout the study timeframe. For the 2013 assessment presented in this chapter, the system recorded roughly 62 million detections in the GP lanes and 95 million detections in the HOT lanes. Traffic volumes are much larger in the general purpose lanes than in the HOT lanes. At any given time, roughly twice the number of tag-equipped vehicles are operating in the GP lanes as in the HOT lanes (i.e. less than 1/3 of tag-equipped users have opted into using the HOT lane). Because HOT lane detectors are located every 1/3 to 1/2 mile, and there are only six GP lane detection stations, detector density in the HOT lanes is higher than in the general purpose lanes (5.8:1 in the southbound direction, 5:1 in the northbound direction), increasing the detections/vehicle/mile in the HOT lane. Only the RFID-equipped vehicles are detected in the GP lanes. Hence, the relative number of detections across the lanes is presented to demonstrate that the number of vehicles and monitored trips involved in the study is very large.

The HOT and GP data originate from the same data source, and these data constitute revealed preference data, which sets this study apart from most previous studies. Although the RFID-equipped vehicles are spread across multiple lanes, the large number of using the lanes means that the RFID-derived speeds, travel times, and other conditions from the vehicles in the GP lanes are representative of actual travel during the peak periods. Speeds differ across GP lanes, and the RFID tags are not uniformly

distributed across these lanes. Nevertheless, the volume of vehicles monitored in the GP lanes is such that the data can be used to represent the conditions of non-HOT travel. Similarly, the data allow for direct comparisons of the conditions between the Express and GP lanes. The process of generating trips and estimating operating conditions from these detections was described previously in the Data Processing chapter.

For this initial study, the data will only include trips that use only a single lane type. Mixed trips (those that include trip segments in both the HOT and GP lanes) are excluded from the analysis until supplemental analyses are undertaken in Chapter 12. Hence, while the following analyses are representative of single-lane-type trip decision making, the behavior of the population subset that uses the lane for only a portion of their trips may differ significantly.

The socio-economic data used in these analyses come from the marketing data described previously in Chapter 3 (Khoeini, 2013; Khoeini, 2014). These data include many demographic variables at the household and individual levels. This chapter makes use of the household level variables, as it is not possible to identify individuals within the observed lane-use data. The household variables used here include income, size, education, and head of household age.

As described previously, the full demographic data set originated from a list of license plates collected by researchers for a previous project at Georgia Tech (Guensler, et al., 2013). These plates were collected quarterly, during peak period hours on I-85, from 2010 to 2012 as part of an HOV-to-HOT conversion analysis. The complete data set included over 300,000 unique license plate records. For the research described here, the subset of the data that could be tied to SRTA data were used. This excluded

corporate accounts, which could not be joined to demographic data. Accounts that joined to multiple households were also excluded. The data were matched via a single-blind process that created a link between observed plates and the privately sourced data without explicitly connecting license plates with registration database data. The results were stored on a secure server at Georgia Tech. A total of 76,764 unique households were paired with the complete SRTA data; 28,953 households were identified within the study timeframe. Previous chapters describe this process and its complications in greater detail.

Variables in the data set included:

- Lane Choice (dependent variable) - HOT lane vs. GP lane (coded as 1,0)
- Trip Length (miles) - Based upon sequential tag reads, irrespective of whether the trip is in the HOT lane or GP lanes
- Toll Amount ($) - Based upon toll paid for HOT lane use or toll that would have been charged based upon GP entry and exit locations
- Trip Direction - Northbound vs. southbound
- HOT Lane Speed (mph) – Space mean speed of trips in HOT lane along the same trip length at the same time
- HOT Transponder Count – Count per fifteen minute bin of the number of tags detected in the HOT lane along the same trip length (surrogate for traffic volume)
- GP Lane Speed (mph) - Space mean speed of trips in GP lane along the same trip length at the same time
- GP Lane Transponder Count – Count per fifteen minute bin of tags detected in GP lanes along the same trip length (surrogate for traffic volume)

- Congested Conditions flag – Indicates speeds less than 40 mph in GP lanes
- Household Income - Demographic data
- Household Size - Demographic data
- Household Education - Demographic data
- Head of Household Age - Demographic data

Figure 107 illustrates the marketing data demographic characteristics of this initial data set in the form of distributions for the 28,953 households. As indicated by the income distribution, more than 40% of the households in the sample have annual incomes between $50,000 and $100,000. The income segmentation categories arose out of the divide illustrated in this distribution. The household size results illustrate that over 30% of the sample households include a single individual, while approximately 40% have two or three vehicles. Roughly 50% of the sample households have members who completed a college education. The proportion of households with graduate degrees is roughly similar to those with only some high school completed. In terms of the head of household age, over 60% are between 35 and 54 years old.

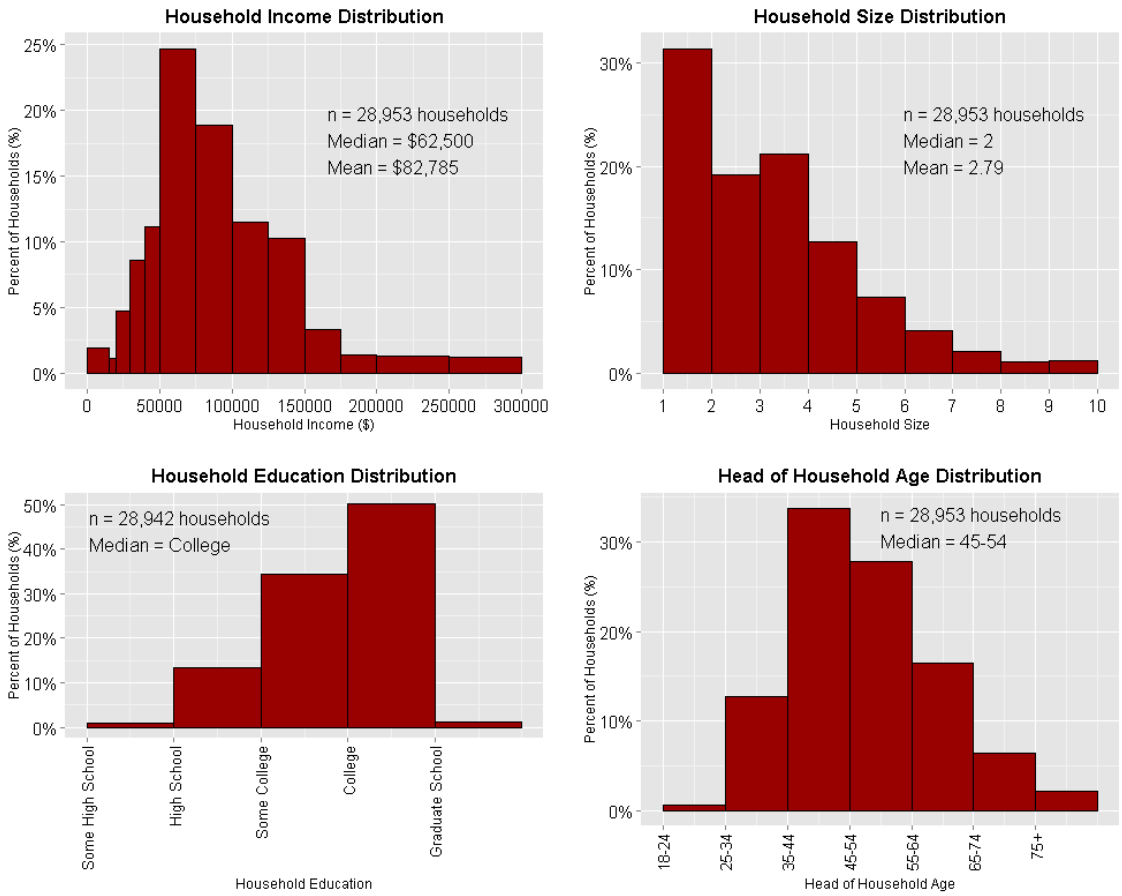


Figure 107: Demographic Distributions of Examined Households

Methodology

The Vehicle detection stream described above delivers disaggregated data from each of the AVI scanners and HOT gantries. Those disaggregate data needed to be processed in different ways for different purposes: vehicles trips were built from the disaggregate detections, average travel times and speeds were calculated for the various start and end locations on the corridor, and total transponder counts were collected for those same locations. The Data Processing chapter earlier in this dissertation provides the details on how these various data sets were constructed and joined together. Briefly, the individual vehicle detections were used to re-construct the trips taken by the Peach Pass holders. The resulting reconstructed trip includes data identifying the RFID transponder, the start

and end time of the trip, the gantry and corridor segment at which the trip started, the gantry and corridor segment at which the trip ended, as well as the overall speed and travel time for the trip. This reconstructed trip data also include those characteristics of the individual GP and HOT portions of trips that use both lane types.

Trips with speeds greater than 100 mph were excluded; this filter was implemented due to detections that were perceived to be mistimed or misreported. Very few of the trips met this criterion; less than 0.1% of trips on any given day. Additionally, trips that had speeds of 0 mph were also removed. This screening eliminated less than 0.1% of the trips on any given day. Finally, trips that started or ended on SR-316 were excluded as there are no General Purpose tag readers on that branch (which made it impossible to compare conditions for the two lane types with the given data).

For each trip, researchers estimated corresponding operational conditions on both lane types. Average trip speeds for the fifteen-minute time interval on the specific day on which the trip was taken were calculated for the segments of the corridor that corresponded to the trip. Similarly, researchers counted all of the distinct Peach Pass transponders that were detected within that road length at the same time. Thus, for each trip, researchers were able to compare average HOT and GP speeds and unique transponder counts along the same road length at the same time. This is not a count of all of the vehicles in the GP lanes, but it is meant to serve as a proxy of such a metric. A congested conditions dummy variable was included for trips that occurred when average speeds were below 40 mph in the GP lanes. The resulting trips were narrowed down to those weekday trips that occurred in the peak hour and direction: from 6:00-10:00 AM southbound, and from 3:00-7:00 PM northbound. Finally, the trips were joined to toll

data to identify the amount charged for trips between the specified origins and destinations at the time of the trip. For trips that occurred in the GP lanes, the toll amount was what the user would have paid had they taken the Express Lanes.

The resulting data set is described below in Table 39. The data set consists of a total of 1,304,079 trips, of which 282,616 were HOT-lane trips and 1,021,463 were GP-lane trips. These trips were extracted from the 2013 data. These trips were taken by 28,953 unique households with corresponding demographic data.

Table 39: Overview of Initial Trip Dataset

	Full Dataset	Lower Income	Medium Income	Higher Income
Households Analyzed	28,953	7,959	12,592	8,402
% of Households by Income	100	27.5%	43.5%	29.0%
Total Trips Monitored	1,304,079	393,069	600,696	310,314
HOT Trips	282,616	80,340	126,745	75,531
GP Trips	1,021,463	312,729	473,951	234,783
% of Total Trips by Income		30.1%	46.1%	23.8%
% of HOT Trips by Income		28.4%	44.9%	26.7%
% of GP Trips by Income		30.6%	46.4%	23.0%
% of Trips in HOT Lane		20.4%	21.1%	24.3%
% of Trips in GP Lanes		79.6%	78.9%	75.7%
Average Trip Speed (mph)	52.3	52.1	52.4	52.6
Average Trip Length (mi)	7.96	7.45	8.14	8.27

Table 39 also shows the proportion of trips taken by different income groups. Here, lower income households were defined as those with incomes less than $50,000; medium income households were defined as those with incomes between $50,000 and $100,000, and higher income households were those with incomes over $100,000. As expected, the number of observed trips for each group was generally in proportion to the size of the income group within the transponder-equipped population. However, the HOT lane usage rates were relatively constant across the three income segments (20% - 24% of their trips). Researchers were surprised by this finding, as it does not generally agree with the existing literature on HOT lanes. This initial data set has a number of

limitations that are discussed later in the chapter, including the fact that the transponders examined here only make up roughly 13% of the active transponder population. Average trip length also noticeably increases across the three income segments, which may be an artifact of geographical clustering along the corridor.

Figure 108 illustrates trip speed distributions by lane type and income segment. The lane type distribution on the left shows the lower variance and greater consistency in Express Lane speeds, which is expected. While the peak Express Lane speed is marginally lower than that of the GP lanes, a higher proportion of the trips are taken at this speed. The GP lanes also see more trips in the lower speeds between 20 and 50 mph. As indicated above, the proportions of toll lane trips by each income segment are not vastly different. It therefore follows that the speed distributions for each income group are very similar.

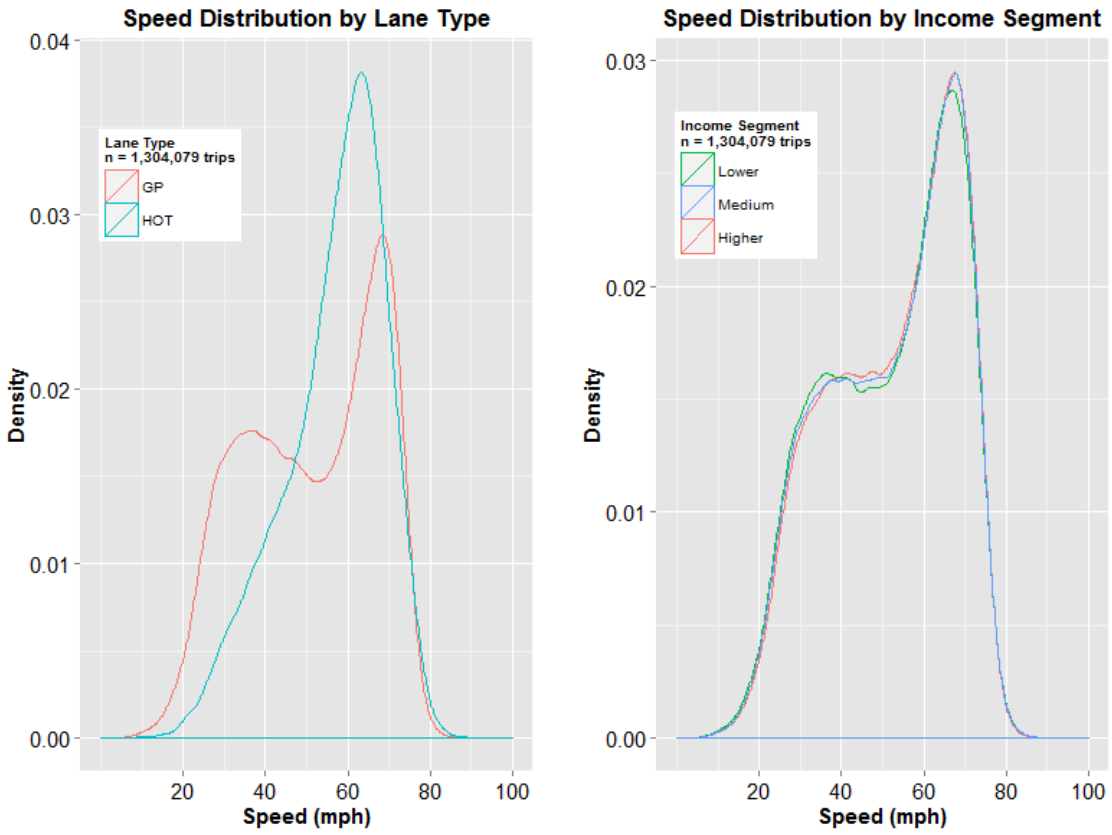


Figure 108: Trip Speed Kernel Densities in Initial Modeling Dataset

In addition to the household income segmentation, k-means clustering on the demographic variables was applied to identify households with similar demographic characteristics. The purpose of the clustering was to reduce potential bias that might be introduced by a dataset containing different numbers of trips for different households. This should help isolate the household-related error component so that it is no longer correlated with other model errors. Table 40 provides an overview of the results of the clustering process. Notable differences include the income and education of cluster two, illustrating highly educated households with higher incomes. The third cluster groups households with older heads-of-household together. Cluster one is the youngest and the

smallest in terms of household size, and has the lowest income. Cluster four has the fewest households and the largest household size.

Table 40: Household Cluster Overview

Cluster	Number of Households	Number of Trips	Mean Household Income	Mean Household Size	Mean Household Education	Mean Household Age
1	11,064	538,492	54,308	1.77	3.36	2.81
2	6,512	245,948	135,881	3.24	3.86	4.22
3	6,483	291,307	59,970	2.09	2.97	4.88
4	4,883	226,878	91,572	5.63	3.29	3.50

Logit Modeling

After processing the data, researchers used binary logit modeling to investigate the potential factors influencing lane choice decisions. The result is akin to a mode choice model, as the HOT Express and GP lanes differ in both price and operating conditions at any given time. The dependent variable was the selected lane type (HOT lane vs.GP lane), with the base alternative set as the GP facility. The average speed, transponder count, and toll amount factors are alternative specific, with generic coefficients. Initial model investigation occurred across the entire sample and involved different demographic variables to examine their effects. The model that was selected and is shown in Table 41 below as the Pooled Model was then used for the segmented and clustered data.

Results of Initial Lane Choice Modeling

Table 41 shows the results of the pooled, segmented, and clustered models, with the t-statistic for each estimated coefficient in parentheses. For the pooled model, all of the coefficients achieved significance at the 95% confidence level. This is not unexpected, as the sample size was very large. A visual inspection of the model results reveals a few

differences across incomes. The estimates of the household income coefficients vary across the three income segments, with the higher income segment exhibiting the largest coefficient. The coefficient for the lower income segment was the only one to fail to achieve significance. This parameter relates to incomes within each segment, however, and may be affected by the actual ranges of incomes within each income segment. Toll amount coefficients are, as expected, negative and significant across all income segments. The household size and education estimators are consistently negative across all income segments. However, it may be that correlations between income and other independent variables are appearing in these results, as the discussion of demographic variable correlation in Chapter 3 (Data Sources) revealed positive correlation coefficients between household income and education and between household income and household size. The clustered models, designed in part to address this issue, are more varied in their estimator magnitudes. Cluster four is the only demographic/market segment for which household size was a positive predictor of HOT lane use; this was the group with the largest household size. Cluster two, with the highest household education and income, had the strongest income effect and the smallest education effect (not surprising given the variable correlation). Only cluster three saw a negative impact from household income; this was the cluster with the highest head of household age.

The clustered models had goodness of fit values similar to the pooled and income-segmented models, with R^2 values ranging from 0.183 to 0.213. The primary purpose of the cluster analysis was to reduce potential bias in previous results from the presence of repeat data. While the resulting fit measures were not largely different, the signs and magnitudes of the coefficients exhibit more variation.

The large amount of data yielded almost universally significant coefficients in each model. However, with very large data sets, a statistically significant individual model parameter does not necessarily mean that the parameter will lead to a practical difference in final model application. Researchers estimated elasticity values for each factor in the model to evaluate the relative impacts of the variables. The disaggregate elasticity values were calculated for each observation and then averaged; with the results shown in Table 42. It should be noted that the repeated observation issue that may bias the models also affects the elasticity results.

Toll elasticity is uniformly inelastic across all income segments and demographic clusters. The values, though small, are the largest of the negative elasticities in the table. Higher income households exhibit higher demand elasticity, approaching unity, with respect to income, while medium income households appear less likely to choose an HOT trip as income group increases. Again, the income parameter reflects the range of income values in that segment. Households from cluster two exhibit the highest sensitivity with respect to income, though the value is still below unitary elasticity. A similar pattern can be observed for household size and age: the higher income segment is more sensitive to both of these factors, but the impact is still small. The lower income segment exhibited the largest response to household education, although all segments and clusters had negative elasticities. The highest sensitivity was observed with regards to trip distance; these values exceed unit elasticity across all segments and clusters. However, this variable was later concluded to be problematic because the fraction of trips that traversed the entire corridor, by definition had to be HOT lane only trips. This issue is discussed further in the Limitations section of this chapter. After trip distance, the most

consistently high elasticity results came from the average speed differences. At approximately 0.35 across all segments and clusters, an increase in HOT lane speed relative to GP lane speed increased the probability of choosing the managed lane by a positive but small amount. Dummy variables (direction, GP congestion) were excluded from the elasticity calculations.

The models presented here do not have high goodness-of-fit values, at least by the McFadden pseudo-R^2 metric. The primary goal of this initial research was not to achieve the best fit, but to compare the results across the different income and demographic segments. These differences between the pooled and segmented models were confirmed with a chi-squared test. The results were significant, yielding a test statistic of 6328, which far exceeded the critical value of 42.3 at the $\alpha = 0.001$ confidence level (again, due to the very large sample size). Similarly, the clustered models were significantly different than the pooled model, with a test statistic well above the $\alpha = 0.001$ threshold.

Table 41: Initial Model Results

	Pooled Model	Lower income	Medium Income	Higher income	Cluster One	Cluster Two	Cluster Three	Cluster Four
Intercept	-5.76 (-122.11)	-4.00 (-39.52)	-2.27 (-9.43)	-17.12 (-70.08)	-4.43 (-55.95)	-17.51 (-69.93)	-3.95 (-38.08)	-5.64 (-37.07)
Average Speed	0.032 (96.1)	0.038 (60.57)	0.031 (63.02)	0.028 (42.38)	0.033 (63.79)	0.031 (40.51)	0.034 (47.84)	0.028 (36.2)
Transponder Count	-0.0026 (-118.37)	-0.0031 (-74.53)	-0.0023 (-73.18)	-0.0024 (-55.82)	-0.0029 (-86.72)	-0.0029 (-57.26)	-0.0020 (-43.17)	-0.0020 (-38.31)
Toll Amount	-0.15 (-90.43)	-0.16 (-49.79)	-0.16 (-63.29)	-0.14 (-41.87)	-0.15 (-54.56)	-0.14 (-35.5)	-0.16 (-43.6)	-0.18 (-45.67)
HOT: GP Congestion	1.90 (244.15)	1.87 (127.19)	1.86 (161.8)	2.03 (130.39)	0.67 (79.96)	0.88 (69.09)	0.56 (48.39)	0.45 (35.49)
HOT: Southbound AM Trip	0.65 (119.84)	0.77 (75.52)	0.55 (70.08)	0.68 (62.5)	0.18 (148.74)	0.18 (96.63)	0.23 (133.81)	0.20 (107.88)
HOT: Trip Distance	0.19 (245.13)	0.21 (141.43)	0.19 (160.85)	0.20 (123.18)	1.87 (154.45)	2.09 (115.2)	1.78 (106.91)	1.94 (103.78)
HOT: log(HH Income)	0.18 (37.9)	0.0082 (0.85)**	-0.15 (-6.72)	1.13 (55.19)	0.062 (8.14)	1.22 (62.63)	-0.045 (-4.69)	0.11 (8.08)
HOT: Household Size	-0.035 (-23.95)	-0.030 (-8.47)	-0.045 (-21.05)	-0.038 (-13.79)	-0.048 (-11.49)	-0.12 (-23.73)	-0.084 (-16.79)	0.022 (5.41)
HOT: Household Education	-0.14 (-40.6)	-0.25 (-42.55)	-0.089 (-16.91)	-0.027 (-3.43)	-0.14 (-25.96)	-0.068 (-5.07)	-0.20 (-26.35)	-0.12 (-13.96)
HOT: Household Age	-0.046 (-21.05)	-0.033 (-8.57)	-0.029 (-9.25)	-0.15 (-28.76)	-0.056 (-9.03)	-0.20 (-29.99)	0.078 (13.2)	0.060 (8.18)
Log-Likelihood	-552750	-157210	-255180	-137190	-228880	-104170	-119090	-96743
McFadden R^2	0.188	0.209	0.175	0.203	0.183	0.213	0.204	0.186
Chi-Squared Test Results (vs. Pooled Model)	N/A	Test Statistic: 6328 Critical Value (0.001): 48.268			Test Statistic: 7729 Critical Value (0.001): 63.87			

Table 42: Initial Models – Elasticity Results

	Pooled Model	Lower income	Medium Income	Higher income	Cluster One	Cluster Two	Cluster Three	Cluster Four
Average Speed	0.35	0.42	0.34	0.30	0.36	0.34	0.38	0.31
Transponder Count	0.086	0.089	0.083	0.087	0.091	0.10	0.073	0.067
Toll Amount	-0.24	-0.24	-0.25	-0.21	-0.22	-0.21	-0.25	-0.30
HOT: Trip Distance	1.10	1.11	1.11	1.14	1.01	1.01	1.31	1.24
HOT: log(HH Income)	0.14	0.0065	-0.12	0.86	0.049	0.94	-0.036	0.084
HOT: Household Size	-0.023	-0.011	-0.030	-0.036	-0.017	-0.11	-0.041	0.029
HOT: Household Education	-0.13	-0.21	-0.086	-0.026	-0.13	-0.070	-0.17	-0.11
HOT: Household Age	-0.045	-0.031	-0.029	-0.15	-0.044	-0.22	0.096	0.058

Limitations of the Initial Modeling Process

While this initial research revealed useful and interesting results, certain limitations must be noted. Mixed trips were excluded from this dataset; only those that occurred in a single lane type were studied. There may be different household behaviors across income groups with respect to partial trip lane use and origin-destination patterns. The use of relative speeds to compare the lane types may have yielded some bias, as free flow speeds do constitute a choice of the drivers at the time they are driving. Inherent for any observed choice in the data set is the individual driver's assessment of the amount of time they believe they will save using the lane, for which no data are available. Even though many different households were observed in the one-year period, the average number of trips per household in the data set was 45.1, and there still may be a significant impact associated with repeat observations of the same users. The standard binomial logit framework used here does not account for repeated observations by the same users, so the results may exhibit bias in that regard. As mentioned earlier, the cluster analysis and segmentation was designed to address this issue. This potential limitation will be addressed again later in the dissertation.

Travel time reliability, which is often cited as a benefit that HOT users are willing to pay for (Brownstone & Small, 2005), is not yet included in this research. In addition, the lack of survey data available for this dissertation meant that this research could not incorporate trip purpose and other attributes that often play a large role in mode choice studies (Li, 2001). The toll rate cap may change user behavior, as the price does not appear to reach market-clearing levels in 2013 (the toll cap is reached and congestion forms on the HOT lane). And while the facility meets its 45mph goal the majority of the

time, congestion still occurs in the lanes. Since the dataset only includes registered RFID tag holders, users without tags are not represented. Finally, the number of transponders examined in this initial study was roughly 44,000, which constituted about 13% of the total active transponder population (approximately 345,000) for which observation data could be paired with demographic data. This may have introduced some significant bias to the results.

The distance variable incorporated into these models was later discarded, as its nature made it highly correlated with the lane choice dependent variable (all through trips on the HOT lane are 15.5 miles in length). The distance variable was calculated by finding the difference between the first and last detection gantry. The issue that made this variable unacceptable for modeling purposes was that the gantry locations are different across the two lane types. As a result, the sets of possible distance values differ across the two lane types. For example, only the Express Lane gantries extend for the entire length of the corridor. So, in cases where the trip distance was approximately fifteen miles, a GP lane use was not possible. Later chapters and models exclude the distance variable for this reason in lieu of segment counts, which do not vary by lane type.

Chapter Overview

The purpose of this initial research was to identify potential factors associated with Express Lane use decisions and examine differences across demographic groups. The data used for these models included vehicle detections and toll amounts from 2013, along with the Epsilon marketing household demographic data. These data were processed and combined to generate HOT and GP lane trips, attributes of those trips and corridor conditions, and socioeconomic attributes of the households making the trips. Binary logit mode choice models were estimated across different income segments and clusters to examine differences in decision making between low, medium, and higher income households and between demographically similar households. The results indicated that the income-segmented models yielded different results than the pooled model at the 95% confidence level, but the parameters were largely consistent across the three segments. The clustered households exhibited more variation in their responses, particularly for the older and larger households. For the year studied, rates of HOT lane use were fairly consistent across the three income groups for which data were available, differing by a maximum of 3.9%. Disaggregate elasticity values revealed low sensitivities to nearly all of the explanatory parameters with the exception of trip distance, and with income among the higher income users. These elasticity values illustrated varying responses to household income and education, for example, across the segmented and clustered households. It is important to note that this was the first stage in the modeling process; a look at the goodness-of-fit measures for the different models indicates that there is a lot of room for improvement. These models also reflect the

limitations of revealed preference data; without accompanying survey data, the results are likely to be less than ideal.

The next chapter will expand the scope of analysis to mixed trips, which occur in both lane types, and to a longer timeframe of data. Future models will incorporate a number of improvements that were not yet included here. The main improvements include testing interaction terms, identifying frequent users (for which lane choice behavior may differ compared to more casual users), incorporating the available panel data, and identifying carpool-mode accounts. The interaction terms will be examined along with the correlation between the existing variables, particularly the demographic variables. The study by Goodall and Smith (2010) found large benefits to modeling frequent and infrequent users separately, which should help in this research as well. As discussed above, panel data methods should reduce the effects of correlation among an individual's repeated choices. Users with carpool-mode account types may make lane choice decisions differently; future models will address these account types. Finally, the behavior of the higher income segment indicates that there may be more variation within that segment; future models will investigate those households at the highest end of the income spectrum more closely.

Certain issues cannot be rectified with the existing data: for example, users may not actually reside at the addresses at which their vehicles are registered (as noted by Granell, (2002)). In addition, there are fewer vehicle detectors in the GP lanes than in the Express Lanes, and those GP detectors are not always adjacent to an HOT detector. This makes it difficult to compare travel times across the lane types directly, which is why this

research used space mean speed for comparison. Limits on the toll rates, which were in effect during the months studied here, may also affect modeling results.

CHAPTER 9

EPSILON-PAIRED VERSUS UNPAIRED TRANSPONDER MODELS

As discussed in previous chapters, the process of pairing SRTA Express Lane use data with Epsilon household demographic data narrows the sample substantially and introduces the potential for bias in analytical results. The purpose of this chapter is to examine the paired and unpaired data sets through basic models using operational factors to investigate the impact of the data loss caused by the pairing process. Throughout this chapter, 'paired' data and 'matched' data refer to those trips which could successfully be joined to the marketing data set. Both the paired and unpaired data include all other joins: travel time, trip stream, etc.

The first section provides an overview of the paired and unpaired data used in the study. The methodology section explains which variables were investigated and the modeling strategy that was employed. The results section then presents and discusses the model outputs. Finally, the chapter addresses the limitations of this analysis and describes the next steps in this research.

Data

The data set used in this analysis consists of the set of all trips in 2013 constructed from the individual vehicle detections provided by SRTA, joined to the additional SRTA data streams described previously in the Data Processing chapter. Those streams include the Trip summary stream, the Toll Rate stream, and the Account stream. The constructed trips are also joined to the travel time and transponder count databases that the researchers created, also using individual vehicle detections. The data set was not joined

to the marketing data; rather the transponders that would have been successfully joined to that data were identified. These matched transponders were those that could be successfully paired to the marketing on February 1, 2013.

Within the 2013 constructed trip dataset, a total of 62,018 (26.5%) transponders were successfully paired with the Epsilon demographic data. The remaining 172,216 (73.5%) transponders could not be paired with demographic data. Table 43 below presents an overview of the two data sets. The data set that is paired with demographic data has roughly 700,000 more trips than the unmatched data set, and yet the unpaired set has 2.78 times the number of transponders as the paired set. Users in the unpaired data set appear less frequently, taking an average of 10.2 trips per transponder, less than a third of the average in the paired data set. Other trip characteristics are more similar: average speed differs by only one mile per hour, and the most frequent start and end segments in Express Lane trips are consistent across both data sets. The unpaired users use the Express Lanes less frequently. Furthermore, the GP-exclusive trip rate of the users in the unpaired data set exceeds that of the paired users by almost 5%. Note that this table includes trips from all days and time periods; it is not restricted to weekday peak-period peak-direction trips.

Table 43: Summary of Paired and Unpaired Data Sets

	Paired Data Set	Unpaired Data Set
Number of Trips	2,471,952	1,748,947
Number of Transponders	62,018	172,216
Average Trips/Transponder	39.9	10.2
Average HOT Trips/Transponder	19.6	4.5
Percent of Transponders with at least one HOT trip	73.8%	34.8%
Average Trip Speed	53.2 mph	54.2 mph
% HOT Trips	14.8%	13.3%
% GP Trips	50.8%	55.7%
% Mixed Trips	34.4%	31.0%
Most frequent HOT entry point – Southbound	Old Peachtree Road	Old Peachtree Road
Most frequent HOT exit point – Southbound	I-285	I-285
Most frequent HOT entry point – Northbound	I-285	I-285
Most frequent HOT exit point – Northbound	Old Peachtree Road	Old Peachtree Road

As mentioned above, Table 43 showed the average number of trips per Peach Pass transponder in the paired and unpaired data sets. To examine this further, Figure 109 illustrates the distributions of the number of trips per transponder. The paired dataset is far less concentrated at the low end, with a more substantial tail approaching the higher trip counts. The unmatched dataset has a much larger proportion of transponders that take only one trip: over 30% higher than the paired dataset. Table 43 shows that a far higher proportion of matched transponders take at least one HOT lane trip (75% vs. 35%). This result agreed with other findings in this dissertation that indicated that the households for which Epsilon demographic data were purchased were more frequent users of the corridor.

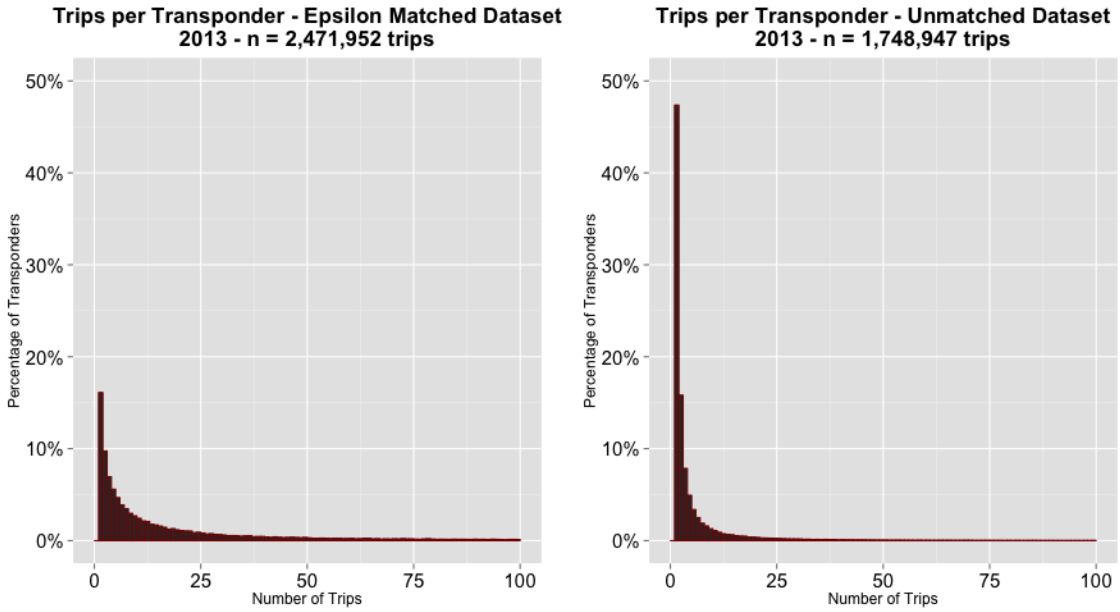


Figure 109: Trips per Transponder - Matched vs Unmatched with Demographic Data

Figure 110 restricts the distributions to the number of toll lane trips per transponder in the matched and unmatched sets. The results are similar to the previous pair of charts, and the differences are similarly stark. In particular, the proportion of users who did not take any Express Lane trips in 2013 is much higher in the unmatched data set (68%) than in the matched data set (32%). Every other trip count bin, from one onwards, is smaller in the unmatched set than the matched set. This is likely associated with households outside of the I-85 commutershed for whom we did not purchase data, such as Georgia State Route 400 users.

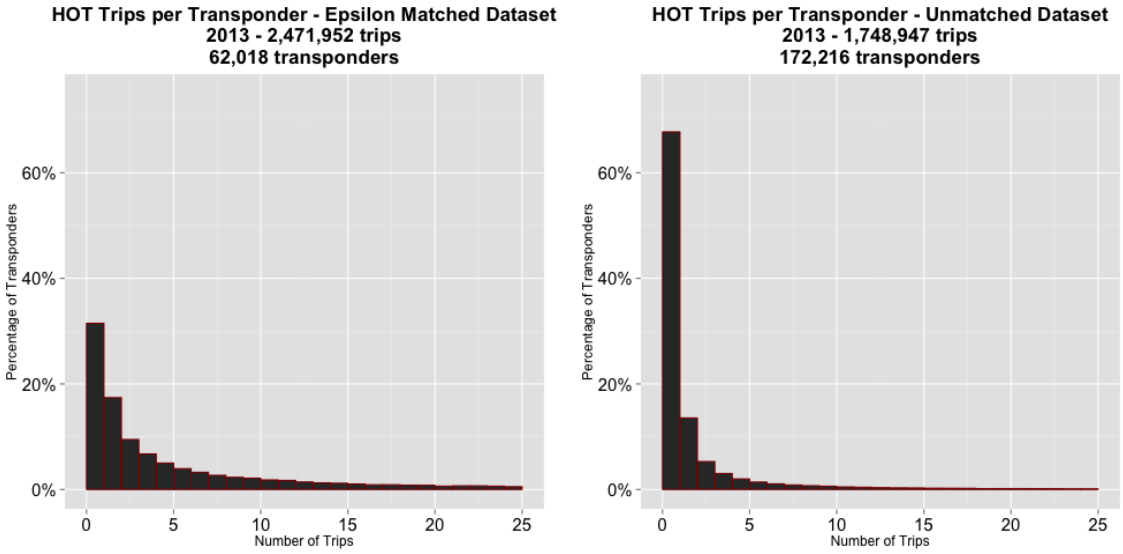


Figure 110: HOT Trips per Transponder - Matched vs Unmatched Data

Figure 111 compares the overall trip speed densities for the two datasets. Here the differences are less noticeable: the unpaired dataset has a slightly higher peak near 65 mph, while the paired trips have marginally more trips near 35 mph. The similarity in distributions is expected given the rates of Express Lane use across the two data sets. The higher peak at higher speeds in the unmatched data set, in which the rate of GP-only trips was higher, may reflect faster GP-lane trips outside of the peak periods.

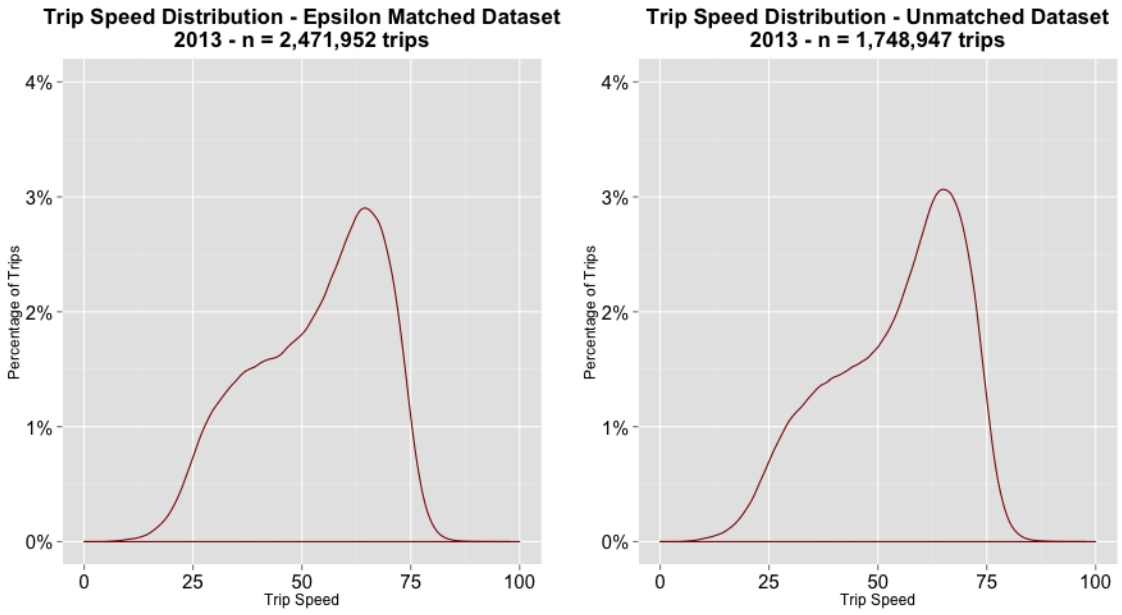


Figure 111: Trip Speeds - Matched vs Unmatched Trips

Methodology

As in the previous chapter, researchers estimated binary logit models using the paired and unpaired data sets. The variables available for the modeling investigation were restricted to operational and trip characteristics; demographic data could not be incorporated as they could not be provided for the unpaired transponders. Variables in the data set included:

- Lane Choice (dependent variable) - HOT lane vs. GP lane
- Toll Amount ($) - Based upon toll paid for HOT lane use or toll that would have been charged based upon GP entry and exit locations
- Trip Direction - Northbound vs. southbound (used for segmentation)
- HOT Lane Speed (mph) – Space mean speed of trips in HOT lane along the same trip length
- HOT Lane density (count) – Count of the number of tags per mile detected in the HOT lane along the same trip segment. Transponder counts are based on 15-minute bins.
- GP Lane Speed (mph) - Space mean speed of trips in GP lane along the same trip length
- Congested Conditions flags – Indicates speeds less than 40, 35, 30, 25, 20, 15, or 10 mph in GP lanes
- Segment dummy variables: Indicate whether the vehicle was detected in each of the five corridor segments (Old Peachtree, Pleasant Hill, Indian Trail, Jimmy Carter, and I-285). This detection can occur in either the Express Lanes or the General Purpose lanes.
- Segment count: Total number of segments in which the vehicle was detected for that trip. The detections can occur in the Express Lanes or the General Purpose lanes.
- Half-hour time interval dummy variables: Indicate which peak-period half-hour interval the trip occurred during. The morning peak period extends from 6:00 AM to 10:00 AM, while the afternoon peak begins at 3:00 PM and ends at 7:00 PM.

- Seasonal dummy variables: Indicate which of the four seasons the trip occurred in. "Winter" includes December, January, and February. "Spring" includes March, April, and May. "Summer" includes June, July, and August. "Fall" includes September, October, and November.
- Day of week dummy variables: Indicate on which day the trip occurred.

As in the initial modeling work, the dependent variable was the choice to use the HOT lane at any point (HOT-incorporating trips vs. GP-exclusive trips), with the base alternative set as the GP facility. The square of the average speed and toll amount factors were alternative specific, with generic coefficients. The first modeling run estimated an intercepts-only model and then generated a series of models looking at each variable in isolation. In the case of multiple dummy variables representing a single factor, such as the start time of the trip, all of the dummies were included. The models restricted the observations to those weekday trips within the peak period hours and directions: southbound trips between 6:00-10:00 AM, and northbound trips between 3:00-7:00 PM. Researchers then estimated multivariate models incorporating all of the previously examined factors to investigate sign changes and other indicators of collinearity among the operation data. After examining each variable on its own, researchers used the random forest technique to evaluate the relative importance of the different factors.

Univariate Paired vs. Unpaired Lane Choice Modeling

The first model estimated included intercept terms only. Table 44 shows the results of this most basic model, with the shares of each sample using the HOT alternative. The intercept values are in line with the share of HOT alternatives; the matched share is slightly higher than the unmatched share in the AM peak models, and the intercept magnitudes reflect this. Similarly, the unmatched users in the afternoon peak choose the HOT alternative less than 50% of the time. The resulting intercept is the only negative parameter.

Table 44: Matched vs Unmatched Models - Intercept Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	0.078***	0.040***	0.152***	-0.047***
	t = 40.801	t = 16.317	t = 79.787	t = -21.237
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.000	0.000	0.000	0.000
Log Likelihood	-758,048.800	-457,673.900	-767,717.900	-558,693.700
LR Test (df = 1)	0.000	0.000	0.000	0.000
Note:				*p **p ***p<0.01

Table 45 shows the results of the paired and unpaired models using the square of the difference in average speeds between the HOT and GP lanes. Here the differences between the paired and unpaired segments were minor, but differences between the morning and afternoon models exhibited behavior that reappears throughout the chapter and the dissertation. The coefficients for speed difference in the afternoon models achieved much higher levels of significance than those of the morning period models; the models' goodness of fit values were higher as well. Within comparable time period models, the results were inconsistent. Matched users exhibited lower sensitivity to lane speed differences in the morning relative to unmatched users, while matched users in the afternoon were more sensitive than their unmatched counterparts.

Table 45: Matched vs Unmatched Models - Speed Difference Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-0.182***	-0.263***	-1.412***	-1.374***
	t = -54.904	t = -62.807	t = -238.213	t = -196.993
I(avgSpeed)	0.020***	0.024***	0.077***	0.064***
	t = 95.760	t = 89.372	t = 279.619	t = 202.597
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.006	0.009	0.057	0.040
Log Likelihood	-753,382.600	-453,582.100	-724,216.900	-536,413.800
LR Test (df = 2)	9,332.425***	8,183.590***	87,001.810***	44,559.790***
Note:				*p **p ***p<0.01

Table 46 presents the results with only toll amount in the model. For trips that did not occur in the Express Lanes, the toll amount used was the toll the user would have paid had they traversed that particular corridor segment at that specific time in the toll lanes. For the general purpose lanes, the toll was always zero. Here we see more of a difference in the morning peak models: higher toll amounts correlated with lower HOT lane use for matched users, but higher HOT lane use for unmatched users. Afternoon peak users also exhibited positive sensitivities to the toll rate. Again, the afternoon model coefficients achieved higher levels of significance.

Table 46: Matched vs Unmatched Models - Toll Amount Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	0.142***	-0.074***	-0.095***	-0.436***
	t = 40.067	t = -16.821	t = -30.916	t = -120.013
tollAmount	-0.022***	0.041***	0.136***	0.215***
	t = -21.424	t = 31.282	t = 100.783	t = 134.348
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.0003	0.001	0.007	0.017
Log Likelihood	-757,819.200	-457,183.600	-762,483.100	-549,175.100
LR Test (df = 2)	459.082***	980.609***	10,469.470***	19,037.170***
Note:				*p **p ***p<0.01

Models in Table 47 include only toll lane density at the time of the trip. The differences between the paired and unpaired models were slight, and while the morning and afternoon peak models had differing signs, the coefficients were all very close to zero. Here the afternoon models do not exhibit the advantage in coefficient significance and goodness of fit that was present in previous models. The models do not explain much, however, and the coefficient significance can be attributed to the large numbers of observations in each of the models.

Table 47: Matched vs Unmatched Models - htDensity Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	0.183***	0.280***	0.001	-0.088***
	t = 55.337	t = 65.289	t = 0.145	t = -22.522
HOT:htDensity	-0.001***	-0.003***	0.002***	0.001***
	t = -38.861	t = -67.882	t = 48.997	t = 12.700
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.001	0.005	0.002	0.0001
Log Likelihood	-757,282.100	-455,225.900	-766,472.400	-558,612.700
LR Test (df = 2)	1,533.320***	4,896.054***	2,490.977***	161.863***
Note:				*p **p ***p<0.01

Table 48 presents the four models with the segmentCount variable, which counted the total number of corridor segments in which the vehicle was detected (in either lane type) for a given trip. This variable replaced the 'distance' variable from earlier models, as that factor was highly dependent on the lane type. The estimated coefficients were positive in all four of the models: longer trips increased the likelihood of Express Lane use. In both the morning and afternoon peaks, the unmatched users saw greater increases in toll lane use probability with increasing trip length than the matched users. The afternoon models again exhibited higher levels of significance in the segmentCount coefficients and better goodness of fit measures overall. While these single-variable models do not have much explanatory power, they do reinforce the idea of separating the morning and afternoon trip models.

Table 48: Matched vs Unmatched Models - Segment Count Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-1.444***	-1.986***	-2.215***	-2.835***
	t = -208.714	t = -223.116	t = -317.788	t = -326.001
HOT:segmentCount	0.411***	0.563***	0.638***	0.757***
	t = 230.199	t = 239.717	t = 357.566	t = 340.554
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.037	0.070	0.097	0.128
Log Likelihood	-729,882.100	-425,493.500	-693,621.400	-487,393.800
LR Test (df = 2)	56,333.390***	64,360.840***	148,192.900***	142,599.700***

Note: *p **p ***p<0.01

Table 49 presents the time of day dummy variables for the trip start time. Start times were aggregated into half-hour increments, with the first increment (6:00 AM in the

morning, 3:00 PM in the afternoon) excluded from the models. In the AM period models, the matched sample coefficient estimates were uniformly larger than those of the unmatched sample. The corresponding t-statistics were larger in all cases as well. The PM peak models followed the same pattern with the exception of the pm1530 coefficient. Taking a trip at that time increased the probability of toll lane use more for unpaired corridor users than for paired users. For the remainder of the estimated afternoon coefficients, that relationship is reversed. The afternoon peak models also exhibit lower goodness of fit measures and t-statistics overall.

Table 49: Matched vs Unmatched Models - Half-Hour Dummies Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-0.588***	-0.410***	-0.270***	-0.406***
	t = -108.509	t = -58.639	t = -43.923	t = -60.979
HOT:am630	0.647***	0.476***		
	t = 87.767	t = 50.003		
HOT:am700	0.791***	0.577***		
	t = 107.600	t = 60.305		
HOT:am730	0.777***	0.620***		
	t = 105.542	t = 63.981		
HOT:am800	0.845***	0.614***		
	t = 111.955	t = 62.228		
HOT:am830	0.880***	0.637***		
	t = 111.121	t = 62.603		
HOT:am900	0.827***	0.457***		
	t = 100.025	t = 44.730		
HOT:am930	0.577***	0.151***		
	t = 66.370	t = 14.353		
HOT:pm1530			0.391***	0.445***
			t = 45.958	t = 48.773
HOT:pm1600			0.559***	0.492***
			t = 68.399	t = 53.944
HOT:pm1630			0.612***	0.482***
			t = 76.470	t = 53.312
HOT:pm1700			0.577***	0.507***
			t = 73.274	t = 56.495
HOT:pm1730			0.472***	0.434***
			t = 59.530	t = 47.502
HOT:pm1800			0.374***	0.308***
			t = 46.019	t = 33.332
HOT:pm1830			0.195***	0.113***
			t = 23.081	t = 11.917

Table 49 Continued

HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.013	0.009	0.006	0.006
Log Likelihood	-747,872.400	-453,703.300	-762,947.100	-555,608.800
LR Test (df = 8)	20,352.840***	7,941.195***	9,541.468***	6,169.773***

Note: *p **p ***p<0.01

The dummy variable for general purpose lane congestion is examined in the models shown in Table 50. The models used the congested40 variable, which is set to one if speeds in the general purpose lane are under 40 miles per hour, because previous modeling exercises have shown its strength relative to the other congestion dummy variables (35 mph down to 10 mph). The results here were similar across the matched and unmatched models: congested GP conditions increased the probability of a decision maker taking a toll lane trip. In the case of afternoon peak trips, paired users exhibited a higher level of sensitivity to general purpose congestion than unpaired users. Again, the afternoon period models exhibit higher t-statistics and goodness-of-fit measures.

Table 50: Matched vs Unmatched Models - GP Congestion Dummy Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-0.298***	-0.339***	-0.480***	-0.580***
	t = -106.654	t = -95.735	t = -172.029	t = -177.298
HOT:congested40	0.725***	0.757***	1.264***	1.067***
	t = 186.475	t = 151.055	t = 316.305	t = 231.152
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.023	0.025	0.068	0.050
Log Likelihood	-740,380.100	-446,060.000	-715,165.400	-531,020.500
LR Test (df = 2)	35,337.350***	23,227.780***	105,104.900***	55,346.280***
Note:				*p **p ***p<0.01

Table 51 shows the results of the seasonality dummy variables on the four models. The effects were relatively consistent across the matched and unmatched models, and across the morning and afternoon peak period models. Taking a trip in spring, summer, or fall increased the likelihood of toll lane use relative to winter trips. Only the matched model for the afternoon peak had a positive intercept value, corresponding to its highest toll lane trip share.

Table 51: Matched vs Unmatched Models - Seasonal Dummies Only

	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-0.033***	-0.092***	0.027***	-0.163***
	t = -8.636	t = -18.369	t = 7.101	t = -35.777
HOT:spring	0.137***	0.144***	0.162***	0.125***
	t = 25.516	t = 20.463	t = 30.364	t = 19.407
HOT:summer	0.119***	0.161***	0.201***	0.155***
	t = 21.878	t = 22.977	t = 37.090	t = 24.291
HOT:fall	0.191***	0.217***	0.135***	0.176***
	t = 34.752	t = 31.121	t = 25.037	t = 27.958
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.001	0.001	0.001	0.001
Log Likelihood	-757,398.600	-457,151.300	-766,932.600	-558,234.400
LR Test (df = 4)	1,300.279***	1,045.336***	1,570.583***	918.432***
Note:				*p **p ***p<0.01

The final univariate models, examining the day of the week on which the trip was taken, are shown in Table 52. The differences between the matched and unmatched sample models were not entirely consistent, though they mostly reflect the higher likelihood of paired users using the Express Lanes. One notable difference between the morning and afternoon peak models was reflected in the coefficients for the Friday dummy variable: in the morning, a Friday trip reduced the probability of using the Express Lanes. In the afternoon, however, Friday trips saw the largest increase in toll lane use probability. The matched and unmatched coefficients for the Friday dummy also had the largest t-statistics among all four models.

Table 52: Matched vs Unmatched Models – Day of Week Dummies Only

	Dependent variable:			
	hotUse			
	Matched Model AM	Unmatched Model AM	Matched Model PM	Unmatched Model PM
	(1)	(2)	(3)	(4)
HOT:(intercept)	-0.037***	-0.060***	-0.013***	-0.198***
	t = -8.485	t = -10.908	t = -3.017	t = -38.702
HOT:tuesday	0.219***	0.206***	0.108***	0.085***
	t = 36.317	t = 26.621	t = 18.208	t = 11.996
HOT:wednesday	0.226***	0.218***	0.146***	0.124***
	t = 37.191	t = 27.860	t = 24.534	t = 17.298
HOT:thursday	0.218***	0.193***	0.244***	0.211***
	t = 35.929	t = 24.730	t = 40.612	t = 29.491
HOT:friday	-0.107***	-0.129***	0.351***	0.330***
	t = -17.448	t = -16.282	t = 56.316	t = 46.162
HOT Share	0.5195	0.51	0.5379	0.4882
Observations	1,094,835	660,476	1,112,188	806,350
R^2	0.003	0.003	0.002	0.002
Log Likelihood	-755,472.400	-456,122.200	-765,842.500	-557,433.600
LR Test (df = 5)	5,152.731***	3,103.383***	3,750.681***	2,520.131***
Note:				*p**p***p<0.01

A final set of models, presented below, incorporated all of the variables that were examined in isolation in the previous models, for the purpose of investigating whether any of them saw changes in their estimated coefficients that may suggest collinearity among the operational variables. Table 53 presents these multivariate models for the morning peak period data sets. The coefficient signs, magnitudes, and t-statistics were largely similar, with the only difference appearing in the am930 time interval dummy variable. The unmatched users were less likely to take a toll lane trip at this time versus the 6:00 AM base interval, all else being equal. The paired users saw a positive

coefficient for this same time interval. The matched model yielded consistently higher t-statistics for each estimated coefficient, though a lower goodness-of-fit value overall.

At the individual variable level, the average speed difference coefficients saw a change in sign from positive to negative when included in the full models. In the univariate models, the toll amount coefficient was negative for the matched AM model and positive for the unmatched model. Here they were negative in both cases. The congested40 dummy variable maintained its sign; collinearity between this variable and the average speed difference may explain the change in the latter coefficient's sign. The season and day of week dummies maintained their signs and relative magnitudes as well. The time of day dummy coefficients were also largely the same, with only the am930 variable in the unmatched sample changing to a negative estimator. This is likely an issue of correlation between the average speeds and transponder counts. This issue is discussed further in Chapter 11 and Appendix A.

Table 53: Matched vs Unmatched AM Models

	Matched Model AM (1)	Unmatched Model AM (2)
HOT:(intercept)	-3.224*** (t = -234.165)	-3.375*** (t = -192.852)
I(avgSpeed)	-0.015*** (t = -54.472)	-0.010*** (t = -28.338)
tollAmount	-0.669*** (t = -327.190)	-0.631*** (t = -239.123)
HOT:congested40	1.234*** (t = 190.193)	1.192*** (t = 140.548)
HOT:spring	0.226*** (t = 37.777)	0.220*** (t = 27.766)
HOT:summer	0.183*** (t = 30.117)	0.222*** (t = 27.771)
HOT:fall	0.464*** (t = 73.811)	0.437*** (t = 54.269)
HOT:htDensity	-0.003*** (t = -58.217)	-0.003*** (t = -49.780)
HOT:segmentCount	0.904*** (t = 329.242)	0.990*** (t = 279.028)
HOT:am630	1.949*** (t = 206.604)	1.706*** (t = 139.562)
HOT:am700	2.316*** (t = 228.233)	2.009*** (t = 153.031)
HOT:am730	2.257*** (t = 216.092)	2.026*** (t = 149.404)
HOT:am800	2.019*** (t = 196.589)	1.741*** (t = 130.162)
HOT:am830	1.689*** (t = 168.221)	1.417*** (t = 108.796)
HOT:am900	1.122*** (t = 114.734)	0.715*** (t = 58.006)
HOT:am930	0.307*** (t = 30.360)	-0.113*** (t = -9.064)
HOT:tuesday	0.238*** (t = 35.408)	0.232*** (t = 26.530)
HOT:wednesday	0.214*** (t = 31.583)	0.210*** (t = 23.784)
HOT:thursday	0.196*** (t = 29.072)	0.166*** (t = 18.846)
HOT:friday	-0.927*** (t = -123.441)	-0.874*** (t = -90.043)
HOT Share	0.5195	0.51
Observations	1,094,835	660,476
R^2	0.155	0.171
Log Likelihood	-640,796.000	-379,319.300
LR Test (df = 20)	234,505.600***	156,709.200***
Note:		*p**p***p<0.01

Table 54 shows the results for the afternoon peak period trips. The average speed coefficients maintained their positive sign in this case, though the previously positive toll amount estimates became negative. The impact of toll lane transponder density remained positive, as did the segment count coefficient. The half-hour interval dummy variables also saw unchanged signs. Coefficient estimates for these dummy variables were lower in the multivariate models until the 6:00 PM time frame, at which point they exceeded the univariate model coefficients. These models continued the trend of marginally-improved goodness-of-fit measures in the afternoon relative to the morning, and among unmatched data relative to matched data. The congested40 dummy coefficients were very similar to those of the earlier univariate models. The largest change appeared in the season coefficients: whereas in isolation the coefficients were uniformly positive, here the spring and summer trips saw a reduced probability of toll lane utilization. The coefficients for fall were much smaller in magnitude and in one case did not achieve significance at the 95% confidence level. The results did not point towards a consistent difference between the paired and unpaired data, though they do indicate that there may be collinearity affecting these models as well.

Table 54: Matched vs Unmatched PM Models

	Matched Model AM	Unmatched Model AM
	(1)	(2)
HOT:(intercept)	-4.843*** (t = -303.926)	-5.547*** (t = -285.030)
I(avgSpeed)	0.014*** (t = 35.205)	0.006*** (t = 13.645)
tollAmount	-0.458*** (t = -202.776)	-0.355*** (t = -134.844)
HOT:congested40	1.272*** (t = 202.514)	1.079*** (t = 146.242)
HOT:spring	-0.207*** (t = -32.856)	-0.194*** (t = -25.510)
HOT:summer	-0.275*** (t = -43.183)	-0.285*** (t = -37.793)
HOT:fall	0.064*** (t = 9.593)	0.001 (t = 0.183)
HOT:htDensity	0.007*** (t = 94.212)	0.010*** (t = 112.416)
HOT:segmentCount	1.107*** (t = 374.951)	1.252*** (t = 342.386)
HOT:pm1530	0.133*** (t = 13.189)	0.248*** (t = 22.793)
HOT:pm1600	0.183*** (t = 18.476)	0.161*** (t = 14.381)
HOT:pm1630	0.320*** (t = 31.659)	0.145*** (t = 12.504)
HOT:pm1700	0.410*** (t = 39.948)	0.286*** (t = 24.091)
HOT:pm1730	0.429*** (t = 41.536)	0.314*** (t = 26.156)
HOT:pm1800	0.499*** (t = 49.091)	0.354*** (t = 30.251)
HOT:pm1830	0.374*** (t = 37.143)	0.260*** (t = 22.780)
HOT:tuesday	0.058*** (t = 8.543)	0.024*** (t = 2.900)
HOT:wednesday	0.090*** (t = 12.988)	0.049*** (t = 5.885)
HOT:thursday	0.124*** (t = 17.657)	0.050*** (t = 5.931)
HOT:friday	0.236*** (t = 32.211)	0.158*** (t = 18.679)
HOT Share	0.5379	0.4882
Observations	1,112,188	806,350
R^2	0.205	0.215
Log Likelihood	-610,606.400	-438,564.300
LR Test (df = 20)	314,222.800***	240,258.700***
Note:		*p**p***p<0.01

Paired versus Unpaired Modeling Discussion

The results of the univariate models did not point to a consistent set of differences between the matched and unmatched corridor users. The matched models yielded larger coefficients than their unmatched counterparts for the time of day and day of week estimators, but lower coefficients for the segment count variable. The toll lane density and speed difference estimators were similar for both samples. The AM period paired user model yielded a negative toll amount coefficient, but this was not the case for the afternoon model. The main differences between the morning and afternoon models in most cases included higher levels of significance for the coefficient estimates in the afternoon and different behavior on Friday trips. Many of the differences between the paired and unpaired models, particularly those involving larger positive coefficients for the paired models, reflected the higher rate of toll lane use among the paired population.

The multivariate models continued this trend of exhibiting no consistent differences between the paired and unpaired models, outside of the rate of toll lane use. Both the AM and PM peak period model sets suggested that collinearity may be affecting the results, particularly in the case of the average speed and season variables.

Random Forest Variable Exploration

To further investigate the impact of the various operational variables on both data sets, researchers used the random forest method of estimating variable importance. Figure 112 shows the random forest variable importance results for the paired and unpaired morning trips. Factors are listed in order of importance (from top to bottom), and the variable's value along the x-axis indicates the impact on model accuracy caused by removing that variable. The results were restricted to the top twenty variables for readability's sake. In all four cases, the simulation sample size was restricted to one half of the full sample so that the computational processes could finish successfully.

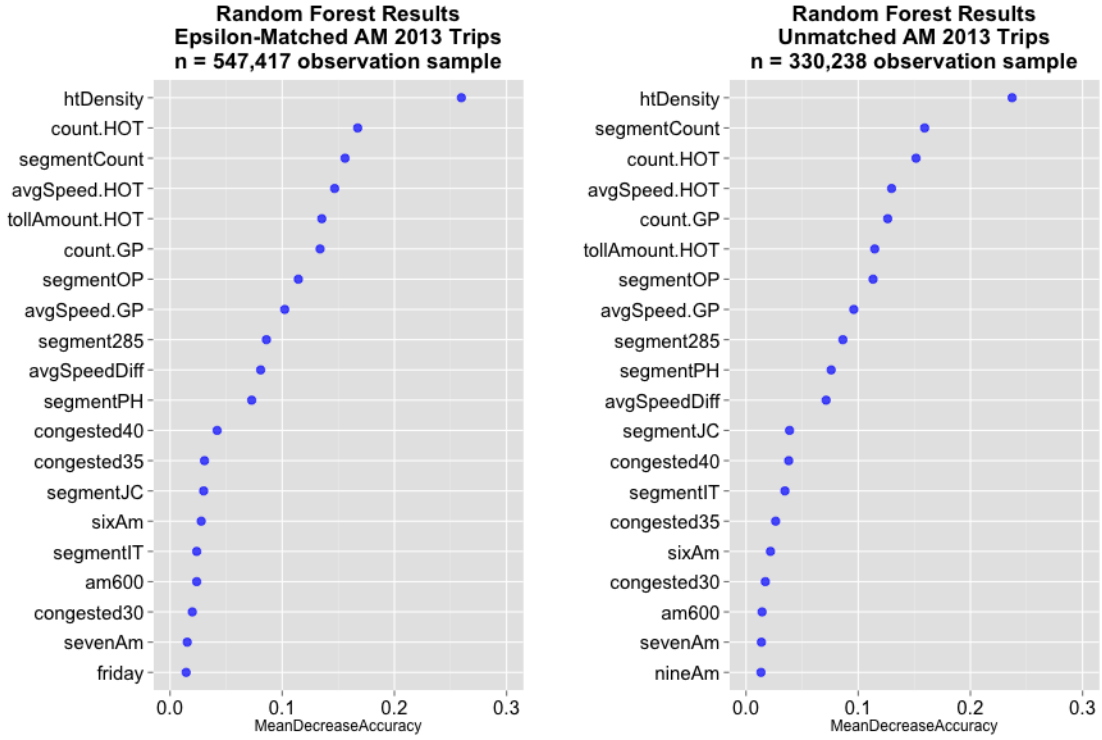


Figure 112: Random Forest Results - Paired and Unpaired AM Trips

The morning trips in the paired and unpaired datasets yielded similar results from the random forest analysis. The top nineteen variables were the same in both cases, with minor differences in order of importance. Only the twentieth-ranked variable differed

between the paired and unpaired samples. Of particular interest was the relative rank of the congested40 variable: in both cases, it sat above all of the other GP congestion dummy variables. Certain variables represented the same underlying data: the htDensity variable used the value of the count.HOT variable divided by the length of the HOT trip. The avgSpeed.GP value was subtracted from the avgSpeed.HOT value to calculate at the avgSpeedDiff variable. The segmentCount variable was the sum of all of the individual corridor segment dummy values (segmentOP, segmentPH, etc.). The presence of these variables, and the similarities among their ranks, reflects the levels of correlation between them. The similarity between the two variable importance charts speaks again to the similarity of the lane-choice decision-making behavior among the matched and unmatched samples.

Figure 113 presents the random forest variable importance results for the afternoon peak period matched and unmatched trips. Here all of the top twenty variables were identical, again with only minor differences in their order. One notable difference in these results versus the morning peak results was the relative position of the avgSpeed.GP and avgSpeed.HOT variables. In the afternoon peak results, the average general purpose lane speed contributed more to model accuracy than the average toll lane speed. The differences, though, were small in magnitude. Again, the congested40 dummy variable outranked the other congested dummies.

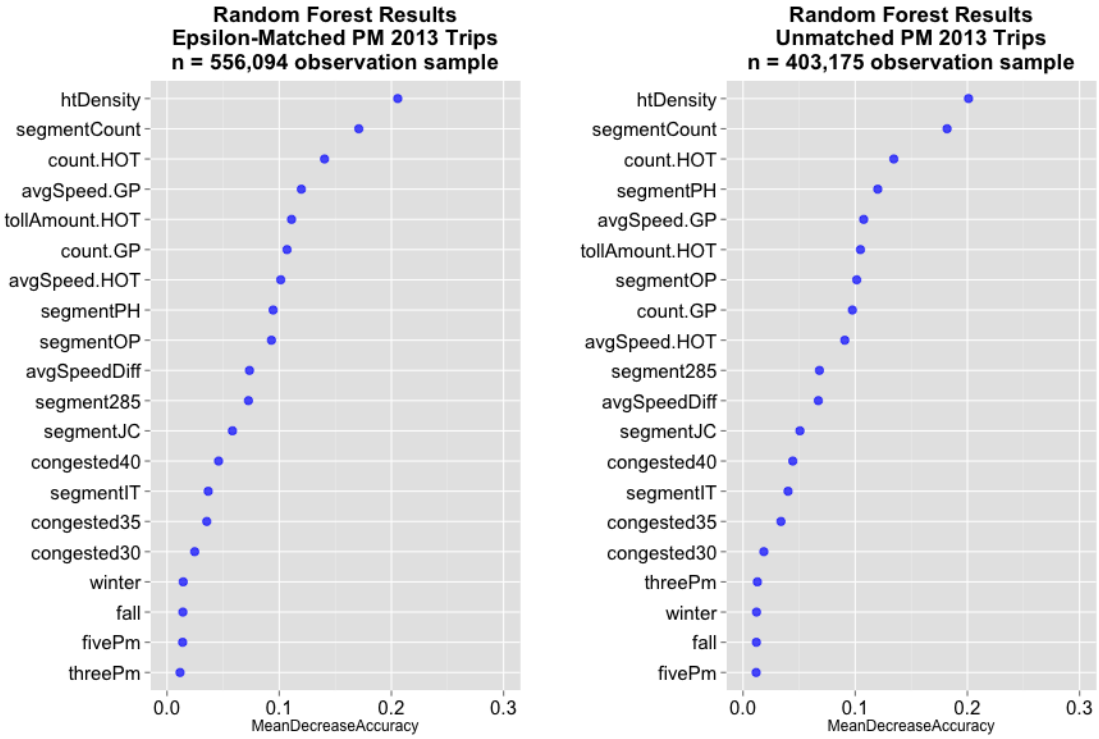


Figure 113: Random Forest Results - Paired and Unpaired PM Trips

Chapter Summary

This exploration of the trip-making behavior of demographic-matched and unmatched households revealed minor differences in the rates of toll lane use between the two populations, but many similarities in the decision-making factors. Corridor users in the paired data set made more corridor trips per transponder, and more Express Lane trips per transponder, than the unpaired users. This was expected as the demographic data purchase was targeted towards the commutershed identified by analyzing the addresses of frequent corridor users (Khoeini, 2014). The examinations of univariate lane choice models revealed no large or consistent differences between the paired and unpaired samples, and the random forest variable importance investigations yielded very similar results for the two populations in both the morning and afternoon peak periods. The combined effect of the various studies in this chapter suggest that while the paired users

were more likely to take corridor and toll lane trips, the factors that influence their decisions were more similar than different.

CHAPTER 10

VALUE OF TRAVEL TIME SAVINGS ANALYSIS

The study of the value I-85 HOT lane users place on the time they have saved by using the facility began with a paper submitted to the Transportation Research Board in August of 2013. That paper was accepted for presentation at the 2014 Annual Meeting and publication in the Transportation Research Record. The research examined travel time savings and toll amounts paid by peak-period HOT users who traversed the entire corridor: Old Peachtree Road to I-285 and vice versa. The TRR paper (Sheikh, 2014) is largely reprinted in the Preliminary Analysis section below. After this initial work, this chapter delves further into the value of travel time savings (VTTS) analysis by expanding the time frame and scope of trips, incorporating methodological changes that allow for partial trips to be examined as well. The chapter uses the paired data set that ties the SRTA lane use data to the Epsilon marketing data to investigate differences in values of travel time savings across income segments. This includes a comparison of the base constructed trip data set with the smaller, processed sample that allows for corridor travel time comparisons. The chapter then examines value of travel time distributions for different income segments and presents the differences among those distributions. The next section presents the results of attempts to fit the value of travel time savings data to various distributions. After that, the chapter presents a comparison of VTTS results for full length trips versus shorter trips. Finally, the chapter concludes with a discussion of the limitations of the analysis.

Preliminary Analysis

As HOT lanes become more prevalent, both in the Atlanta metropolitan region and across the country, an understanding of the way users respond to the lanes and the benefits they derive is important. Such an understanding can inform future implementations, increasing their efficiency and the welfare gains of the customers. In that spirit, this research uses data from the I-85 Express Lanes to investigate users' value of travel time savings and willingness to pay distributions. This avenue of investigation is common to HOT lanes as the results can be used to help design pricing algorithms that satisfy throughput and revenue goals. The results may be useful for other cities that are designing HOT lanes, and for the extensions of the system that are under consideration in Atlanta.

In addition to comparing overall HOT and general purpose (GP) lane performance, this research also examines willingness to pay vs. frequency of facility use. The travel time and reliability measures are compared for infrequent users, frequent users who use the Express Lanes between two and three times a week, and very frequent users who use the HOT lane at least three times a week. The Express Lanes are also contrasted with the leftmost GP lane to generate a more conservative estimate of I-85 Express Lane travel time savings. Finally, the study compares the total value of time saved by HOT users to the time-value using the average wage rate in the Atlanta metropolitan region.

Data Description

The first source of data used in this analysis was the individual vehicle detection data stream provided by SRTA and discussed earlier in Chapter 3, Data Sources. The 35 detectors in the Express Lanes and 13 detectors in the general purpose lanes allow the SRTA system to detect vehicles with Peach Pass RFID transponders in both lane types. The resulting data stream provides a unique transaction number, the unique identification number associated with the detected transponder, the specific lane in which the vehicle was detected, the gantry at which the detection occurred, and the timestamp of the detection. These data are transmitted to Georgia Tech on a daily basis.

The second source of data used in this analysis was the Express Lane trip summary stream. These data provide trip characteristics for all trips in the HOT lane on a daily basis. Characteristics include start and end times, start and end points, whether the trip was in 'TOLL' or 'NON-TOLL' mode, the toll amount paid, and the transponder identification number. 'NON-TOLL' trips are those taken by vehicles with HOV3+, emergency vehicle, and toll exempt accounts (such as alternative fuel vehicles). Toll amounts can be zero when the operating agency overrides the dynamic system, such as in the event of an incident. Unlike the RFID detection stream, these data present only a single record for each Express Lane trip. Only trips in 'TOLL' mode that had toll amounts greater than zero were included in this analysis.

The time frame for this preliminary analysis was September, 2012 through May, 2013 (nine months). While the HOT facility opened in October of 2011, technical issues prevented the use of general purpose detection data necessary for this analysis until August of 2012. Within the nine months under examination, approximately 100 million

vehicle detections and 3 million Express Lane trips were recorded. This study focuses on the weekday peak periods of 6-10 AM (southbound) and 3-7 PM (northbound).

The initial study of corridor through-trips reported in this section calculated travel times for the I-85 corridor from the aforementioned vehicle detection stream. To identify trips by vehicles that traversed the entire corridor, the paper examined records of vehicles that were detected at the northern- and southern-most general purpose lane detectors, and vehicles that were detected at the northern- and southern-most Express Lane detectors, on the same day. The study examined only these trips that traversed the entire length of the corridor. One significant note is that due to the lack of general purpose lane detectors on the SR-316 segment of the corridor, trips between SR-316 and I-285 were not considered in this analysis. Those trips account for 8.6% of all toll lane trips in 2012, compared to 10.9% for trips between Old Peachtree and I-285.

These general purpose readers do not span the entire length of the corridor: the Express Lanes extend approximately two miles beyond the range covered by the general purpose detectors. The resulting corridor length examined in this research was 13.5 miles, or approximately 88% of the total corridor. The corridor travel time was calculated by calculating the difference between the timestamps of the two detections. These records were separated into detections in the general purpose and toll lanes, as well as in the northbound and southbound detections.

Travel Time Filtering

The method for calculating travel times described above introduced a number of possible issues. Detections at the endpoints of the corridor did not guarantee that the vehicle traversed the length of the corridor in a single trip. Users chaining trips may have left the interstate to make a stop (e.g. food or gas), only to return much later, yielding unusually long travel times. In addition, this method did not protect against mixed trips, in which the user traveled in both the HOT and general purpose lanes. As long as she or he started and finished the journey in the same lane type, the intermediate portion of the trip was not automatically considered.

To control for this potentially confounding data, researchers implemented filters to remove certain records. The first of these filters was a limit of two hours for the overall travel time. This value, selected arbitrarily, was judged to be a reasonable 'first-pass' method of eliminating detections that might have been the result of separate trips, such as one in the morning and a second in the afternoon. The second filter addressed the mixed trip issue. Researchers identified vehicles that were detected traveling in the same direction in both the HOT and the general purpose lanes on the same day and then removed these trips from the data set for this initial analysis. The remaining trips were undertaken by vehicles detected in only the HOT lane or only the GP lanes. Finally, researchers implemented a filter consisting employing travel time mean and standard deviation calculations. This filter maintained a running average, as well as a running standard deviation, of thirty travel times. Records that had travel times that were within two times the mean and three times the standard deviation were maintained, while the others were removed. This filter helped remove trips that exited the freeway to make a

stop and then returned later. The researchers implemented this filter to more precisely identify and remove chained trips, as well as those data corresponding to multiple trips in the same direction on the same day. These filters were applied to all of the chronological travel times from September 1, 2012 through May 31, 2013.

As discussed above, the general purpose detection gantries do not cover the entire span of the Express Lanes. To address this, the trip summary data, which contains the toll paid for each journey, was joined to the RFID detection data so that the disaggregated detections for each trip could be identified. This join of the trip summary and RFID detection data allowed researchers to calculate travel times through the span of the corridor that was covered by the general purpose detectors, and to connect those times to the toll paid for the entire trip. Because the general purpose detectors cover 88.1% of the corridor length, that proportion of the toll amount was used in the successive calculations. A potential limitation of this study is that this factor assumes uniform congestion between the monitored 88% and unmonitored 12% of the I-85 corridor.

The nine months in this preliminary study produced a total of 151,517 trip summary records in the southbound direction and 176,725 in the northbound direction. The RFID detection data set contained 141,143 southbound HOT travel times, 376,654 southbound general purpose travel times, 145,886 northbound HOT travel times, and 274,852 northbound general purpose travel times. After joining the trip summary records to the RFID tag read data stream, 108,411 southbound trips and 104,786 northbound trips remained for analysis. These trips contained both positive toll data and HOT travel time data for the subsection of the corridor that could be directly compared to the general purpose lanes.

Reliability Calculations

The RFID detection dataset was then used to calculate average general purpose travel times and travel time reliability for both lane types. The general purpose travel times were aggregated into daily fifteen-minute harmonic means. Travel time reliability was evaluated in fifteen-minute bins at the monthly level using the buffer index measure. The formula, provided by the Strategic Highway Research Program, is as follows:

$$Buffer\ Index = \frac{95th\ percentile\ travel\ time - Average\ travel\ time}{Average\ travel\ time}$$ (Margiotta, 2013)

The buffer index was applied to the average general purpose travel times and to the individual HOT trip times to generate buffer times. The buffer times were then added to the trip times to compute the planning time metric, where planning time is defined by the Federal Highway Administration (FHWA) as the "total time a traveler should allow to ensure on-time arrival" (FHWA, 2006). The results for the AM and PM peak periods can be seen below in Figure 114 and Figure 115.

Value of Travel Time Savings Calculations

To examine travel time savings in the HOT lane, the HOT toll and RFID detection dataset was joined to the set of average travel times for the general purpose lanes. This allowed for the direct comparison of actual Express Lane travel times to average general purpose travel times on that same day by 15-minute time intervals. The average number of general purpose corridor trips used to compute these averages was 30 in the southbound direction and 25 in the northbound direction. Trips in which the HOT travel time was higher than the GP time, and thus the VTTS was negative, were removed from the analysis in the TRB paper. This was justified by the SRTA policy to refund the tolls paid by users who experience breakdown conditions in the HOT lane. Those trips are

retained elsewhere in this dissertation in the choice-based models of Chapters 8 and 12. Additionally, trips in which the time savings was less than five seconds were excluded in this preliminary analysis as they were not significantly different than zero. Similar research by Wood and Burris (2014) also removed trips with negative or very low values of travel time savings. These parameters resulted in the removal of 186 (0.31%) observations from the northbound trips and 808 (1.47%) observations from the southbound trips. Note that further publications of this research will involve a sensitivity analysis of this parameters for exclusion.

Researchers also calculated general purpose lane average travel times for vehicles that used only the leftmost general purpose lane (GP lane one). These vehicles were detected in the left GP lane at each scanning gantry, though they could potentially have changed lanes and returned in between general purpose lane detection points. The resulting dataset contained individual Express Lane trip records along with the time saved relative to average GP travel times.

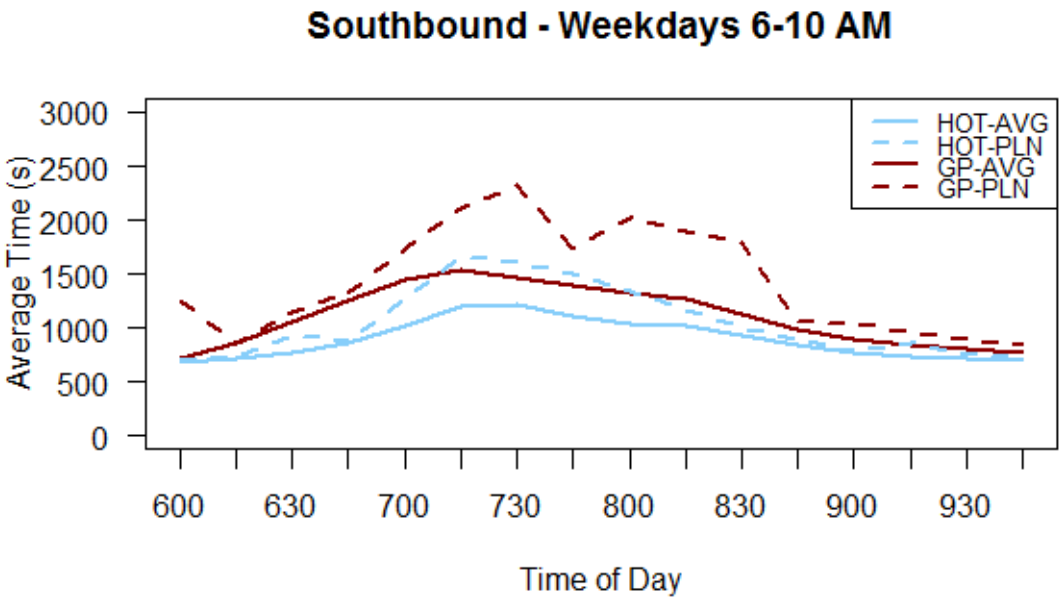
Travel Time Variability and Frequent User Groups

A similar procedure connected the HOT travel time data with the planning time results in both the Express and general purpose lanes, allowing for the difference in planning times to be computed. In the absence of survey data, it remains unknown whether the users are expecting increased reliability in the HOT lane, rather than only travel time savings. The study also identified frequent and very frequent users of the Express Lanes to investigate the differences in their use of the HOT lanes. 'Frequent' users were defined as those who use the lane at least twice a week, or seventy-five times over the nine month interval. 'Very frequent' users were those who used the lane at least 115 times, or roughly three

times a week (selected arbitrarily to begin examining differences in lane use). The study developed separate value of travel time savings for the frequent, very frequent, and infrequent users to see if differences existed. Finally, researchers compared the total toll paid by all Express Lane users in this dataset to the value of that time using the average wage rate in the Atlanta region.

Average Travel Time and Planning Time Results

The average morning peak southbound travel time from Old Peachtree Road to I-285 was 889 seconds in the HOT lane and 1047 seconds in the general purpose lanes. In the northbound direction, the average travel times during the afternoon peak were 798 seconds in the HOT lane and 976 seconds in the general purpose lanes. Figure 114 illustrates how those times, along with the associated planning times calculated from the buffer index, vary across the morning and afternoon peaks. The I-85 Express Lanes provide substantial travel time and reliability benefits, especially in the morning peak period. Figure 114 also indicates that travel times are lower and more consistent in the afternoon peak, but the Express Lanes still provide travel time and reliability improvements. For the figure below, the buffer time values were calculated across all nine months of peak period data.



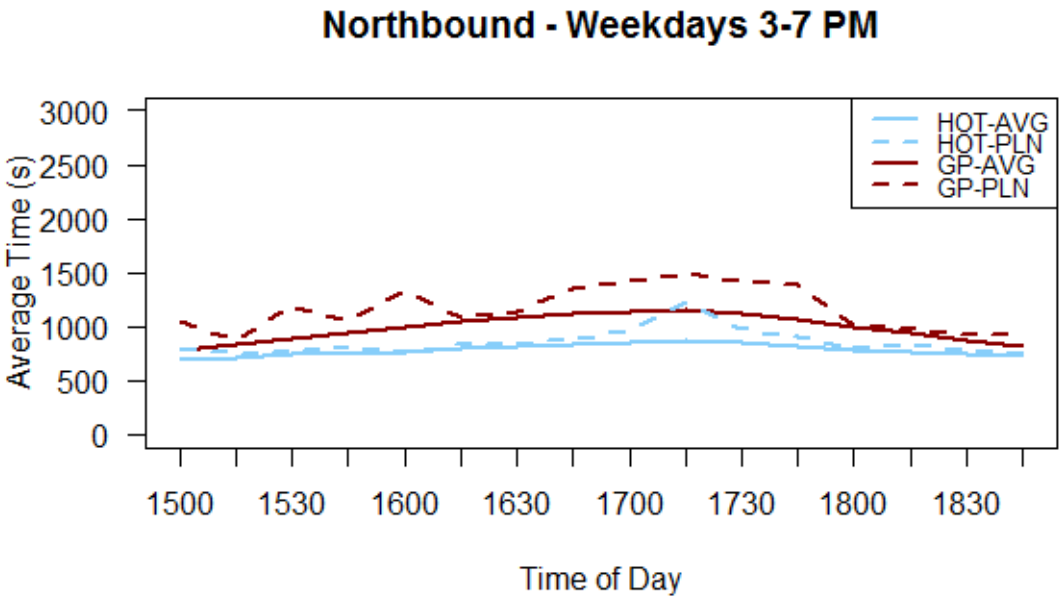
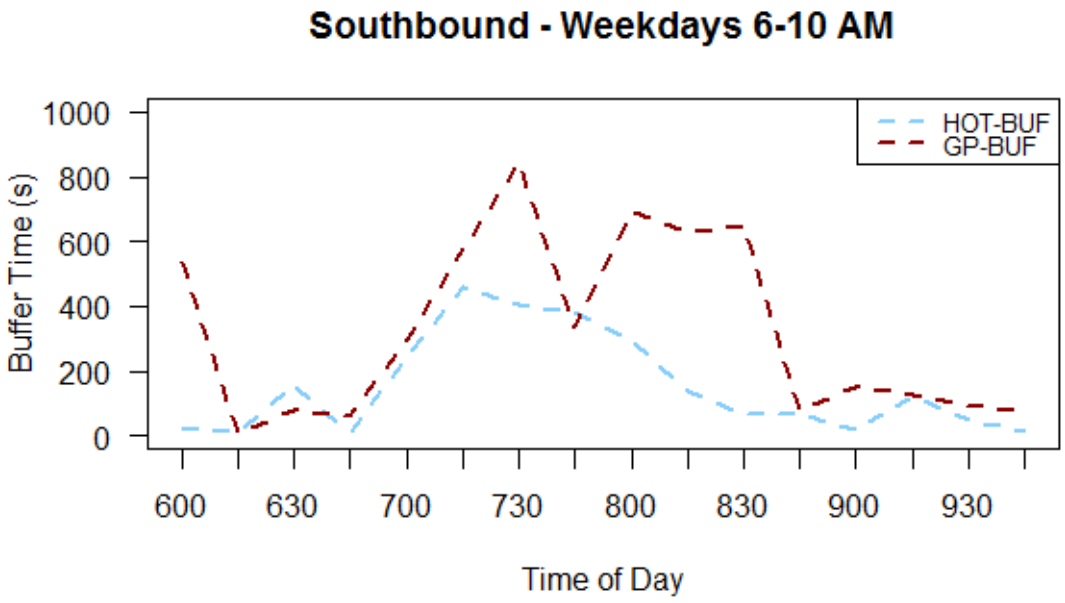


Figure 114: Preliminary Analysis - Average Travel and Planning Times

Buffer Time Difference Results

Buffer time, as defined by the FHWA, "represents the extra time (or time cushion) that travelers must add to their average travel time when planning trips to ensure an on-time arrival" (FHWA, 2006). The buffer time values for each fifteen-minute interval, calculated across all nine months, are shown in Figure 115 for both the morning and afternoon peak periods. The morning peak sees similar results for the HOT and general purpose lanes, with the HOT lanes reporting a marginally higher buffer time than the GP lanes for two of the time intervals. The relatively low buffer times for the 7:45 – 8:00 AM period do not indicate an improvement in traffic but rather show that this period is very consistently congested. Express Lane reliability is more consistent in the northbound direction, and the GP buffer time figures are lower as well. Again, the HOT buffer time marginally exceeds that of the GP lanes for two of the time bins.



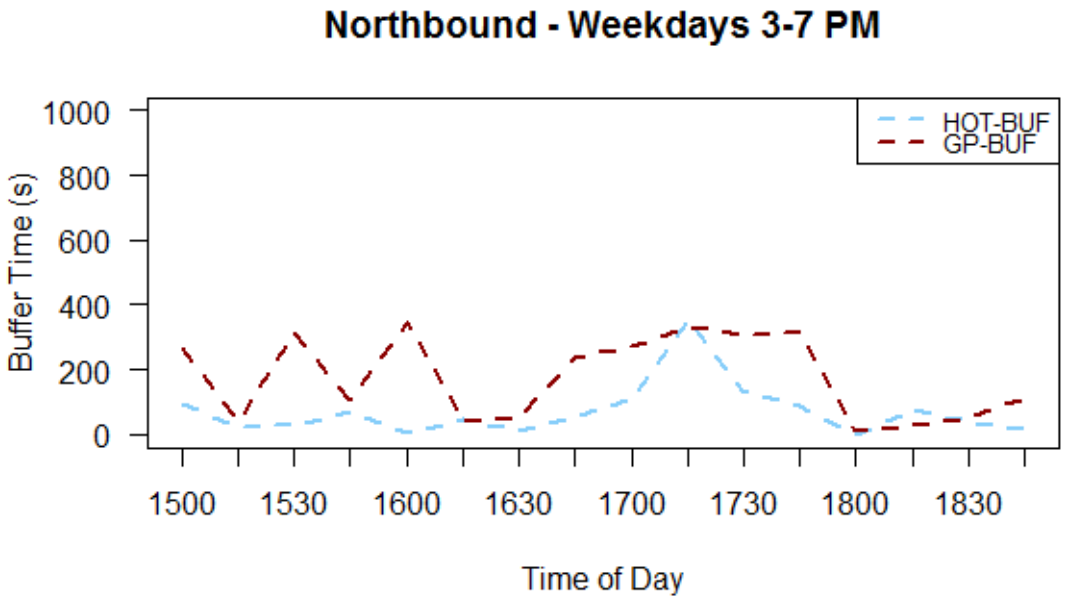


Figure 115: Preliminary Analysis - Average Peak Period Buffer Times

Value of Travel Time Savings Distributions

For each Express Lane trip, the study used the toll paid (multiplied by the trip length reduction factor of 0.8806 discussed earlier) and compared the travel time to the average GP travel time for that specific day, hour, and fifteen-minute interval. Using the toll amount and the travel time difference, the user's value of time saved was calculated for the HOT lanes. The range of values of travel time savings was from $0.20/hour to $4,000/hour for the study corridor over the nine-month study period. Note again that these figures are based on the five-second minimum travel time difference and the exclusion of negative values of travel time savings. The low values in this range occurred when the operating agency set the toll to off-peak rates, including values as low as $0.05. This may have been due to an incident in the lane. The high values result from trips in which the travel time difference equaled the cutoff value for time savings of five seconds. In both the southbound and northbound plots in Figure 116 below, the tail of the distribution stretches far beyond the limit of the chart. The resulting distributions resemble gamma distributions, with the southbound figure yielding higher mean and median values ($55/hour and $36/hour respectively) and more dispersion than the northbound (mean of $34/hour, median of $26/hour). A later section in this chapter explores fitting these data to the gamma and other distributions.

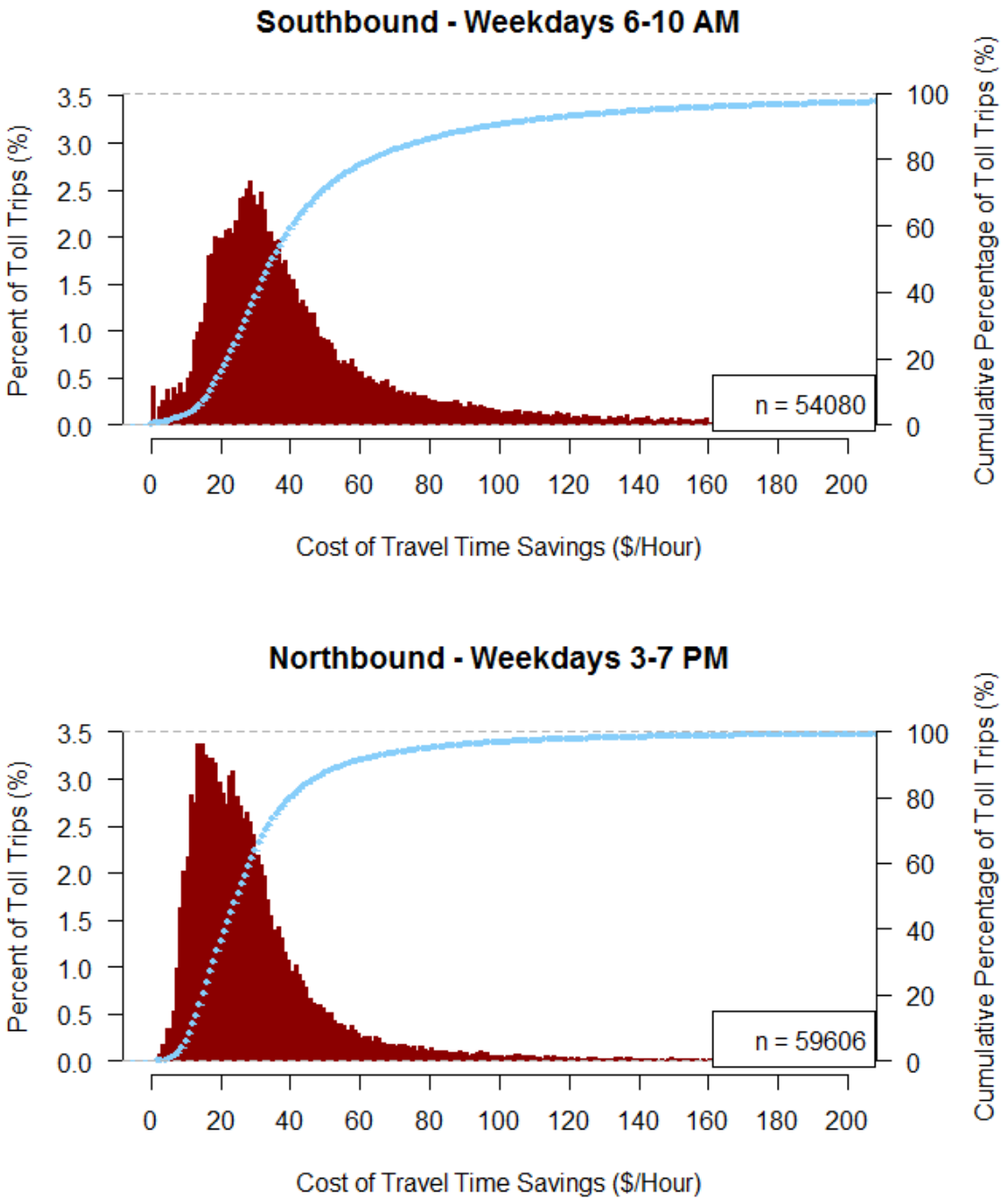


Figure 116: Preliminary Value of Travel Time Savings Distributions

Summary of Value of Travel Time Measures

Table 55 presents the differences in median measures of travel time savings, planning time savings, and values of travel time savings for infrequent users, frequent users, and very frequent users in the preliminary analysis. The 25^{th}- and 75^{th}-percentile values are given in the square brackets. These values were calculated using average travel times across all general purpose lanes, as well as average travel times for users of GP lane one (the left-most lane). This preliminary study also examined the leftmost lane to find a more conservative estimate of HOT lane benefits. The operating assumption was that the left lane would see lower travel times than the combination of all GP lanes, which is generally true for the full corridor traverse, and thus the resulting HOT benefits would be lower as well. Isolating the left lane restricted the number of trips used to calculate the average travel times: across all lanes, an average of 30 southbound trips and 25 northbound trips were used for each time interval, while a mean of 7.7 southbound and 1.7 northbound trips were averaged for each left-most lane interval.

The results in Table 55 indicate a number of interesting and sometimes contradictory trends. Express Lane users tend to save more time in the morning peak than in the afternoon, although the amount of time saved is reduced when compared to only the leftmost GP lane. This holds true across all user groups. One interesting observation from the table is that frequent users tend to save less time per trip on average than infrequent users. Relative to all GP lanes, the infrequent user group saves the most time in both directions. Hence, infrequent users may only be using the lane when it provides greater benefits. The very frequent user group may be using the lane irrespective of toll levels and travel time savings, whereas the infrequent user group

appears more judicious in their choice to use the lane. However, infrequent users no longer save the most travel time when all are compared to GP lane one users. The planning time savings are also greater in the morning than in the afternoon. The median value of travel time saved per mile was 28.85 seconds/mile in the southbound direction and 23.40 seconds/mile in the northbound direction. These values were estimated using the study length of 13.72 miles.

Median values of travel time savings were higher in the morning than in the afternoon. The infrequent users demonstrate the lowest VTTS compared to the other toll lane frequency user groups; this relationship holds in both the AM and PM periods. Looking at the Values of Planning Time Savings reveals contradictory patterns: more frequent users demonstrate higher VPTS values when examining all general purpose lanes, but this relationship is not maintained when looking at GP lane one. In addition, the VPTS figures relative to GP lane one are much higher than those relative to all lanes. This makes sense as the time savings are lower when the HOT lane is compared to GP lane one, while the tolls remain the same.

Table 55: Preliminary Value of Time Calculations (50th, 25th, and 75th percentiles)

Measure	Southbound – AM Peak		Northbound – PM Peak	
	All GP Lanes – Median [25%, 75%]	GP Lane 1 – Median [25%, 75%]	All GP Lanes – Median [25%, 75%]	GP Lane 1 – Median [25%, 75%]
Travel Time Saved per Trip (seconds) by All HOT Users	396 [187, 629] (n = 54,080)	370 [171, 593] (n = 51,059)	321 [200, 477] (n = 59,606)	267 [158, 408] (n = 21,617)
Travel Time Saved per Trip (seconds) by Infrequent Users (<75 trips)	405 [194, 636] (n = 31,794)	377 [175, 599] (n = 29,949)	328 [202, 486] (n = 37,404)	267 [158, 411] (n = 13,242)
Travel Time Saved per Trip (seconds) by Frequent Users (>=75 & < 115 trips)	367 [163, 604] (n = 6,064)	340 [144, 571] (n = 5,733)	312 [200, 465] (n = 8,464)	273 [163, 411] (n = 3,172)
Travel Time Saved per trip (seconds) by Very Frequent Users (>= 115 trips)	388 [188, 622] (n = 16,222)	367 [174, 590] (n = 15,377)	306 [194, 461] (n = 13,738)	261 [153, 398] (n = 5,203)
Planning Time Saved per Trip (seconds) vs. GP Lanes	499 [263, 792] (n = 50,158)	361 [167, 578] (n = 42,602)	425 [273, 600] (n = 59,501)	265 [158, 398] (n = 21,040)
Toll ($)	$4.35 [$2.20, $5.59] (n = 54,080)		$2.02 [$1.62, $2.72] (n = 59,606)	
Toll ($) Infrequent Users (<75 trips)	$4.35 [$2.20, $5.59] (n = 31,794)		$2.02 [$1.62,$2.59] (n = 37,404)	
Toll ($) Frequent Users (>=75 & < 115 trips)	$3.43 [$2.06, $5.50] (n = 6,064)		$2.06 [$1.62,$2.59] (n = 8,464)	
Toll ($) Very Frequent Users (>= 115 trips)	$4.35 [$2.33, 5.59] (n = 16,222)		$2.06 [$1.62, 2.72] (n = 13,738)	
VTTS All HOT Users ($/hour)	$36.04 [$25.39, $56.02] (n = 54,080)	$39.08 [$27.01, $62.71] (n = 51,059)	$25.66 [$17.36, $37.16] (n = 59,606)	$31.49 [$21.77, $49.05] (n = 21,617)
VTTS ($/hour) Infrequent Users (<75 trips)	$35.51 [$24.84, 54.83] (n = 31,794)	$38.63 [$26.50, $61.38] (n = 29,949)	$24.95 [$16.73, $36.59] (n = 37,404)	$31.07 [$21.27, $48.35] (n = 13,242)
VTTS ($/hour) Frequent Users (>=75 & < 115 trips)	$36.93 [$25.81, $58.42] (n = 6,064)	$39.86 [$27.70, $66.37] (n = 5,733)	$26.54 [$18.34, $37.18] (n = 8,464)	$31.23 [$22.08, $48.82] (n = 3,172)
VTTS ($/hour) – Very Frequent Users (>= 115 trips)	$36.78 [$26.27, $57.31] (n = 16,222)	$39.52 [$27.78, $63.81] (n = 15,377)	$26.96 [$18.64, $38.72] (n = 13,738)	$32.57 [$22.75, $50.62] (n = 5,203)

Burris reported a median value of travel time savings of $73/hour for 6:00 to 10:00 AM for I-394 in Minneapolis in 2008. The median value reported here for the same morning hours is lower ($36/hour), but one important note is that the travel time savings in Minneapolis were low: "The small difference between GP and HOT-lane speeds resulted in very small TTS. Thirty-five percent of travelers on the MnPass lanes paid for an average TTS of less than a minute." The difference was starker in the afternoon peak: in Minneapolis, the median VTTS was $116/hour from 2:00 to 7:00 p.m., much higher than the $26/hour reported here. The VTTS results that Burris reported from San Diego were also higher than those found in this study, with median values of $49/hour in the morning and $54/hour in the afternoon. Time savings were also lower in those cases, with median morning and afternoon values of 1.16 minutes and 1.11 minutes respectively (Burris, 2012). Finally, the median toll amounts are similar across all of the frequency groups with one exception: frequent morning users have a median value nearly a dollar less than the other user groups. Their interquartile range remains similar, however.

Value of Time Saved by the I-85 Express Lanes

Finally, this preliminary study compared the total value of the time saved by Express Lane users to an independent average value of that time for the Atlanta metropolitan region. For this comparison, an average value of time of $22.80 per hour was used, which is the average hourly wage rate for the Atlanta metropolitan region as reported by the US Bureau of Labor Statistics in 2012 (U.S. Bureau of Labor Statistics, 2012). In the southbound direction, HOT users valued the total travel time they saved more than the average Atlanta resident would, based upon wage rates. This was true relative to all GP

lanes and to GP lane one, though the difference was greater with GP lane one. The northbound HOT users were much closer to the average Atlanta worker in their value of the travel time saved, with a difference of only $225 across all trips. Relative to GP lane one, that difference increased to over $11,000. These differences reflect the higher speeds in the leftmost lane, which resulted in lower time savings for HOT users. Recent work by Khoeini (2013) on the household incomes of HOT users and non-users supports this finding: the average household incomes of HOT users exceeded those of non-users by over $10,000 per year.

Table 56: Preliminary Value of Time Saved Findings

Measure	Southbound AM Peak All GP Lanes	Southbound AM Peak GP Lane 1	Northbound PM Peak All GP Lanes	Northbound PM Peak GP Lane 1
Total Travel Time Saved (hours)	6,532.09 (n = 54,080)	5,832.79 (n = 51,059)	5,970.98 (n = 59,606)	1,719.44 (n = 21,617)
Sum Tolls Paid ($)	$233,682.50	$224,873.30	$155,195.50	$57,729.77
Sum of Apportioned (88.06%) Tolls Paid ($)	$205,503.00 (n = 54,080)	$197,760.00 (n = 51,059)	$136,363.40 (n = 59,606)	$50,726.56 (n = 21,617)
Value of Travel Time Saved – Average Atlanta VOT ($)	$148,931.70	$132,987.50	$136,138.30	$39,203.23

Discussion and Limitations of Preliminary Analysis

This research examined willingness-to-pay distributions for users of the I-85 Express Lanes, and compared users who use the lane infrequently to those who use it two or three times a week. The median value of travel time savings figures fell within the range of values seen in the literature, but were lower than those reported by a similar study (Burris, 2012). Results for infrequent users indicated higher levels of travel time savings and lower VTTS figures for that group relative to all general purpose lanes. These users may be more selective in their lane choice, paying for trips only when the benefits are higher than average.

This preliminary analysis also compared the travel time variability of the HOT lanes relative to that of the GP lanes and found reliability benefits in the Express Lanes. It is possible that HOT lane users expect these reliability benefits when they make the choice to pay for the lanes, but this effect could not be isolated by revealed preference data alone. In the absence of survey data, the values of travel time savings and potential

reliability benefits could not be separated. The literature concerning this subject, discussed previously in the Literature Review chapter, indicated values of reliability ranging from roughly $1/hour to $20/hour. The VTTS results in this study fell at the upper end of the range suggested by the econometric literature; more data are required to see if the same holds true for the value of reliability. The total value of the time saved by the HOT users exceeded the value of that time using Atlanta's average wage rate.

These preliminary findings contribute to the understanding of HOT lane users and the benefits they derive from their paid trips. As could be expected, these users value their time more highly than the average Atlanta resident. Furthermore, the results illustrate potential differences between users in the Atlanta metropolitan region and those in other cities. Using VTTS results from other HOT implementations may result in sub-optimal throughput or revenue results on potential new HOT facilities around Atlanta. The remainder of this chapter focuses on expanding the data and methods used in this study. One limitation of this work was the sample of trips: the corridor-length journeys investigated here comprise approximately 11% of all the HOT trips. Excluding the segment of the facility along SR-316, which does not have general purpose vehicle detectors, may also have biased the results. Focusing on trips that used only the HOT or GP lanes further narrowed the scope of this preliminary analysis. The use of survey data would address a significant limitation of the revealed preference data set. Survey data could be used to separate users' values of travel time savings and reliability benefits, as well as provide trip purpose data to better understand willingness-to-pay differences.

A more comprehensive analysis, starting in the next section, will include trips that traverse shorter segments of the HOT lanes, and trips that use both lane types throughout

their duration. It will also include trips that begin at different points along the corridor, and not just the northern- and southern-most locations. In addition, this preliminary analysis used only nine months of data out of the three years that are used elsewhere in this dissertation. Finally, this research did not incorporate the Epsilon marketing data to look at demographic differences; this will also be addressed later in this chapter.

Expansion of Analysis

The limited analysis discussed above was performed for the purposes of a Transportation Research Board (TRB) paper and poster presentation. This dissertation expands on that analysis in a number of ways: by including a longer timeframe (three years' worth of data) and by including partial corridor trips (those that enter mid-facility or depart the facility along the corridor), as well as those trips that use both the HOT and GP lanes during the trip. Whereas the previous analysis examined only nine months of trips in 2012, this analysis now examines all of calendar years 2012, 2013, and 2014. In addition, this section uses the Epsilon marketing data to compare results across different income segments and illustrates the differences in those distributions. The chapter then describes attempts to fit the VTTS data to various distributions. The chapter ends by comparing VTTS results for full length trips versus shorter trips and then discussing the limitations of the analysis.

Methodological Changes

This expansion required a modification of the methodology used in the TRB paper. That paper included full-corridor trips only, which started at the first gantry and ended at the last. The inclusion of partial or 'mixed' trips complicated the analysis as there were now many more possibilities for the trip start and end locations. This issue created the most immediate impact in the calculation of the travel time saved. Because the previous analysis included only one pair of HOT lane gantries and one pair of GP lane gantries, comparing the travel times across the two lane types was straightforward. While the gantries were not perfectly aligned along the corridor (as discussed previously, the GP gantry length was 88% of the HOT gantry length), it was easy to apply a linear reduction factor to the HOT travel times. This factor carried the assumption that congestion was comparable in the excluded 12% of the corridor.

Expanding the analysis to include partial trips brought greater complications due to the mismatched alignment of the HOT and GP gantries. That is, there are very few instances where an HOT gantry is directly adjacent to a GP gantry. As such, the segment lengths between gantries are not uniform across both lane types. The distance between the first and last HOT gantry in the Old Peachtree Road segment is not the same as the distance between the two GP lane gantries in the same segment. In fact, the majority of segments include only one GP scanner (spanning all GP lanes) within the segment. Figure 117 below illustrates the difference in gantry numbers and locations across the two lane types in the Old Peachtree Road section of the corridor. The section contains three green GP-lane scanners, labeled SCAN-N6, SCAN-N7, and SCAN-S1. There are nine HOT-lane scanners, labeled OP01 through OP09.

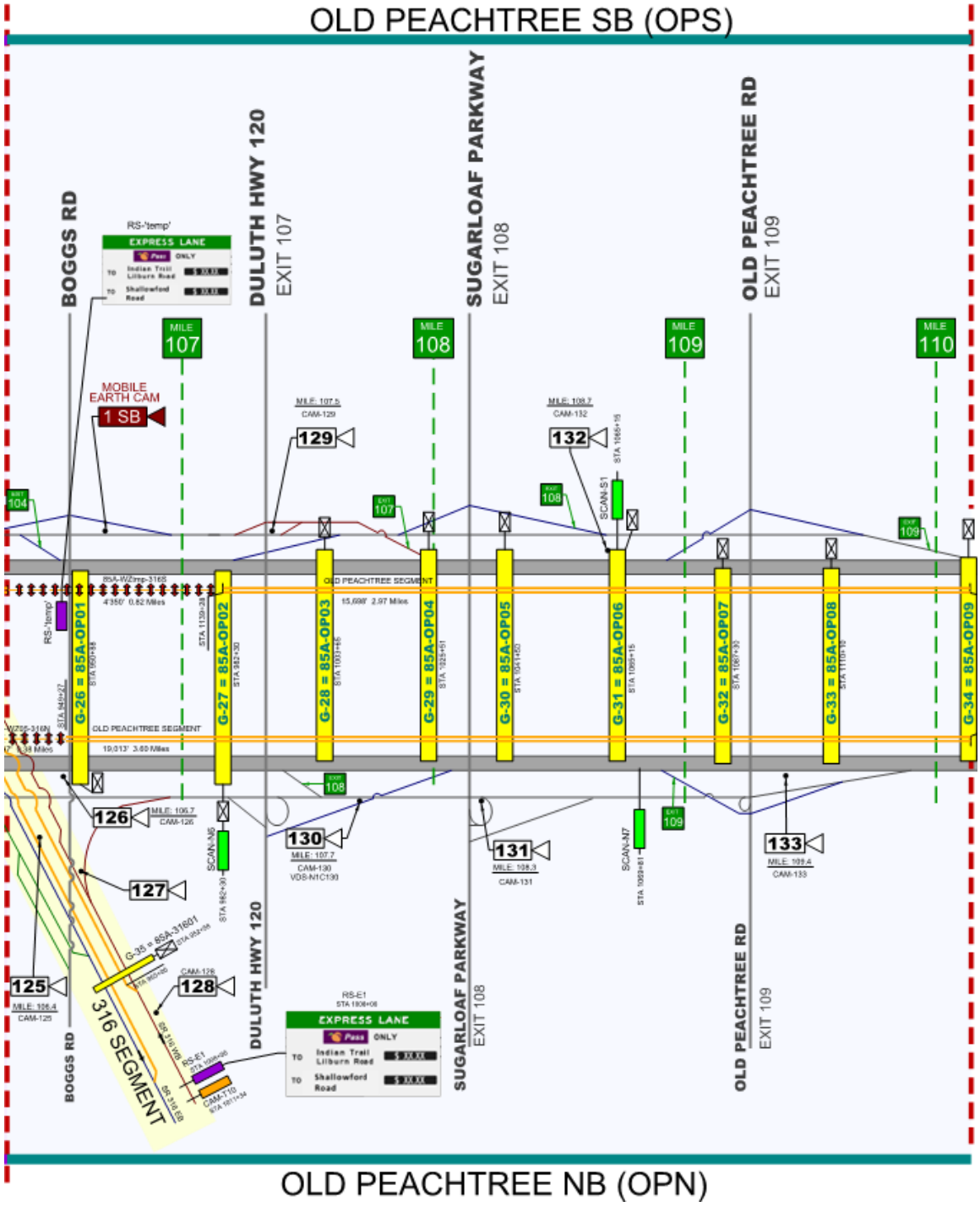


Figure 117: I-85 Express Lanes Straight-Line Diagram

To address the differences in gantry placement, average GP lane speeds in adjacent segments were used to estimate the time saved by the Express Lanes. This allows researchers to control for the differences in segment length caused by the mismatched gantries, but also creates its own issues. The assumption that congestion is uniform along the segments in question is still present. Similarly, varying congestion levels in the portion of a segment contained in one lane type but not the other will affect the average travel time across that segment. The estimates of travel time savings were then converted to value of travel time savings by dividing the toll amount paid by the time saved in hours, so the resulting figure for each trip was reported in $/hour (dollars paid per hour saved). This method resulted in very long tails at the positive end of the spectrum and potentially high values at the negative end, resulting from trips in which the Express Lanes did not deliver travel time advantages over the general purpose lanes. For the purposes of this chapter, namely comparing results across household income segments and attempting to fit the resulting distributions, trips with value of travel time savings results under $0/hour and over $500 per hour were excluded from the data set.

The expansion also generated its own issues and limitations, including the loss of data that occurred when joining constructed corridor trips to average travel time calculations. This issue is addressed in more detail in the Data Pairing and Join Loss section of the Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter of this dissertation. There are a number of reasons that a constructed trip may not join with an average HOT lane or GP lane travel time value. There may have been insufficient users at the time to compute average travel times for that segment of the corridor, though this should be less of an issue when examining peak period trips. It may

also be that the trip includes mistimed vehicle detections, such as a later gantry reporting a detection before a gantry that precedes it physically.

These mistimed detections plagued the RFID readers for the first few months of operations; by March of 2012 they had significantly reduced in number. In the joining scripts, these trips with mistimed detections often report physically impossible start and end gantries. For example, a trip may 'start' at the fifth southbound GP gantry and 'end' at the second. Unless the vehicle was driving northbound in the southbound lanes, this is an error. The number of constructed trip misdetections per month is illustrated below in Figure 118; readers can see the drop in misdetections resulting in a low, stable level by March of 2012. This issue is also discussed in Chapter 6, Data Quality and Treatment.

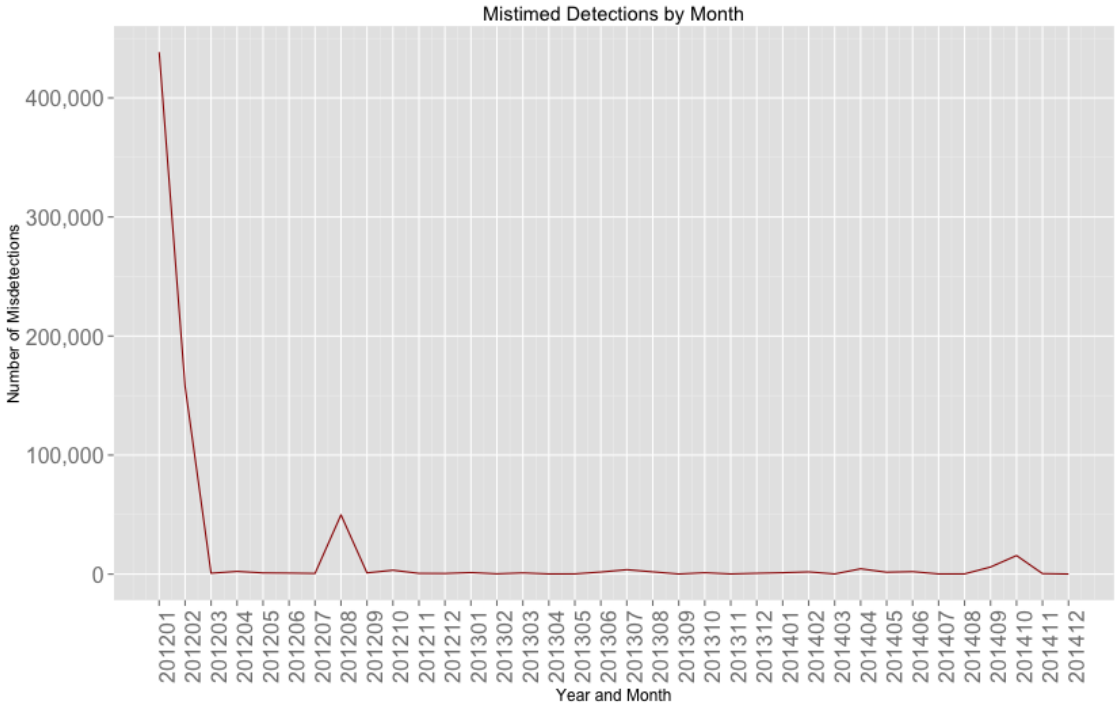


Figure 118: Mistimed Detections in Vehicle Detection Data

In addition, any trip that started or ended on SR-316 could not be joined to an average travel time, as GP lane travel times could not be calculated for that segment of

the corridor. As mentioned earlier, SR-316 has HOT but no GP vehicle detectors and so no trips or travel times starting or ending on that road can be constructed or computed. In 2013, 12.18% of Express Lane trips started on SR-316 and 8.65% of Express Lane trips ended on SR-316.

Trips also may not have been successfully joined to travel times if no travel times were recorded for that time interval and corridor length. This may occur during off-peak hours, for example, when use of the HOT lane is limited. Finally, a speed filter applied to the HOT and GP trip speeds eliminated data as well. This filter removed trips whose speeds were 0 mph or greater than 100 mph in either the HOT or GP portions of the trip. These values often occurred as a result of the mistimed detections reported earlier.

Comparing Travel-Time-Joined Trips to All Constructed Trips

The Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter of this dissertation outlines the amount of data preserved at each stage in the joining process, in the Data Pairing and Join Loss section. The data loss table in that section indicates that in January of 2013, 43.1% of the full constructed trip data set was represented in the travel time-joined data set used in this analysis. The remaining trips were excluded at various stages in the joining process; the majority of these exclusions occurred as a result of the Epsilon demographic data join. Of interest to researchers is the question of the nature of these excluded trips; whether they were randomly distributed or whether bias was present in their exclusion. In the case of SR-316 trips, it was clearly a case of bias as all of the trips starting or ending there were removed. Table 57 presents a comparison of included and excluded trips from January 2013 for the purposes of identifying other biases.

Table 57: Overview of Constructed Trips versus Travel Time-Joined Trips

	All Constructed Trips	Travel Time-Joined Trips
Number of Trips	1,076,511	464,847
Number of Transponders	120,822	40,957
Average Trips/Transponder	8.91	11.35
Average HOT Trips/Transponder	3.05	3.57
Percent of Transponders with at least one HOT trip	41.5%	51.7%
Average Trip Speed (mph)	63.8	62.6
% HOT Trips	12.4%	9.7%
% GP Trips	65.8%	68.5%
% Mixed Trips	21.8%	21.8%

The most striking difference between the two data sets is the difference in the number of trips and transponders represented. The travel-time-joined-trips make up 43.1% of the full constructed trip data set for January 2013, and 33.9% of the total transponder count appears in this more restricted sample. In most other aspects, the differences are less pronounced. Users in the joined data set take more trips and marginally more toll lane trips per transponder, though their overall rate of toll lane use is 2.7% lower than the full sample. The proportion of transponders in the narrower sample with at least one instance of Express Lane use is a full 10% higher than that of the full constructed trip data set. Figure 119 and Figure 120 go into greater detail with regards to trip counts by showing the distributions of trips per transponder and toll lane trips per transponder, respectively, for January 2013.

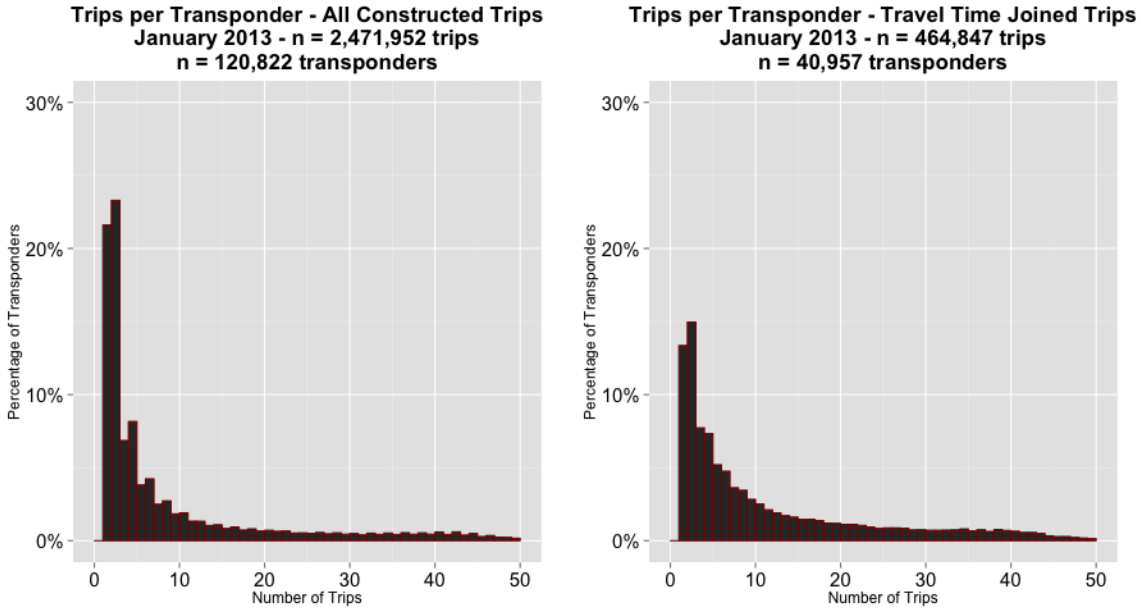


Figure 119: Trips per Transponder - All Trips versus Travel Time Joined Trips

The distributions in Figure 119 reveal that in both the full constructed trip set and the travel time joined trip set, more transponders have two associated trips than one. Besides this observation, it is also notable that the travel time joined transponders have more representation among the higher trip counts. In this case, transponders that remain by the time the travel time join has occurred are more frequent corridor trip-takers.

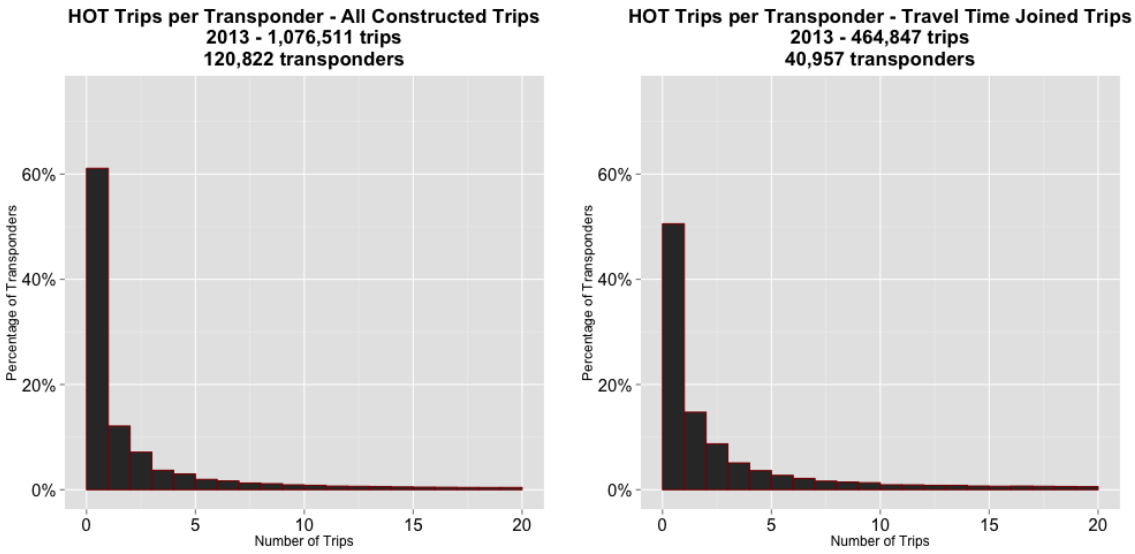


Figure 120: HOT Trips per Transponder - All Trips versus Travel Time Joined Trips

Figure 120 presents the trip distributions for the toll lane trips only. As expected from Table 57 above, the transponders in the narrower travel time sample are more likely to have used the Express Lanes at least once in the month of January, 2013. The remaining differences, between the rates of users with two or more HOT trips, are much smaller. Overall, the travel time-joined sample includes fewer but more frequent users of the corridor in general and of the Express Lanes in particular.

Demographic Component of Value of Travel Time Savings

A further expansion of the value of travel time savings analysis involved the demographic data provided by the Epsilon data purchase. These data allowed researchers to join the trip data to household demographics, so that these results could be compared for different demographic segments. Specifically, this section compares the resulting value of travel time segments across users in differing household income segments. As in the Initial HOT Use Choice Analysis chapter of this dissertation, the income segments are defined based on annual household income: households in the Lower segment earn

less than $50,000 per year, households in the Middle segment earn between $50,000-100,000 per year, and households in the Higher segment earn more than $100,000 per year.. The distributions presented below are for 2013; the results for 2012 and 2014 can be found in the Appendix. An overview of the results is provided at the end of the section in Table 58.

2013 Southbound Distributions and Differences

This section presents the southbound peak periods' value of travel time savings distributions for calendar year 2013. Figure 121 illustrates the distribution of value of travel time savings for users in the lower income segment. This sample consists of 9,692 transponders from 6,086 unique demographic-matched households, making 162,013 trips. The median value of $42.55 is $6.51 higher than the $36.04 value reported for southbound users relative to all GP lanes in Table 55 above. While this may be attributed to the change in the time frame under examination, the narrowing of the sample by the Epsilon marketing data join is a potential source of sample bias as discussed in the Limitations section of this chapter and in Chapter 7.

The addition of mixed trips of both lane types also changes the resulting distribution: in the southbound direction, those trips typically have higher mean and median values of travel time savings than the unmixed corridor trips. Those trips are explored later in this chapter as well. The effect of these mixed trips thus complements the narrowing of the set of users in the sample. The maximum allowable toll rate increased in 2013 as well, which may potentially push the distributions to the right if the time saved remains constant. Again, if that is occurring, its effects appear to be supplementing the sample narrowing effects. The 95% confidence interval around the

median value in Figure 121 was estimated through bootstrap analysis: 1000 simulations consisting of 1000 observations arrived at the upper and lower bounds. That confidence interval is portrayed by the pink shaded region on the figure.

Figure 122 illustrates the medium income segment. This distribution reflects 260,922 trips, 14,991 transponders, and 9,577 households. Here the median value of $41.18 is $5.14 higher than that of the initial analysis (which included only through trips on the corridor) as shown earlier in Table 55. Interestingly, the median value of the medium income segment is over $1 lower than that of the lower income segment. The bounds of the bootstrapped 95% confidence interval are also lower than those of the lower income segment by roughly $1 each. Otherwise, the two distributions appear strikingly similar.

Figure 123 presents the higher income segment. This includes 157,743 trips, 9,396 transponders, and 6,020 households. The median value here of $41.80 is again higher than that of the full set of users over nine months in the initial analysis. A notable characteristic of the 2013 distributions is that the medium income segment exhibits the lowest measures of centrality among the three samples; also striking is the fact that the lowest income segment has the highest mean and median values. The differences among all three income segments are small: the largest difference in median values, between the lower and medium income segments, is only $1.37. Those two segments also exhibit the largest difference in mean VTTS values: $2.14.

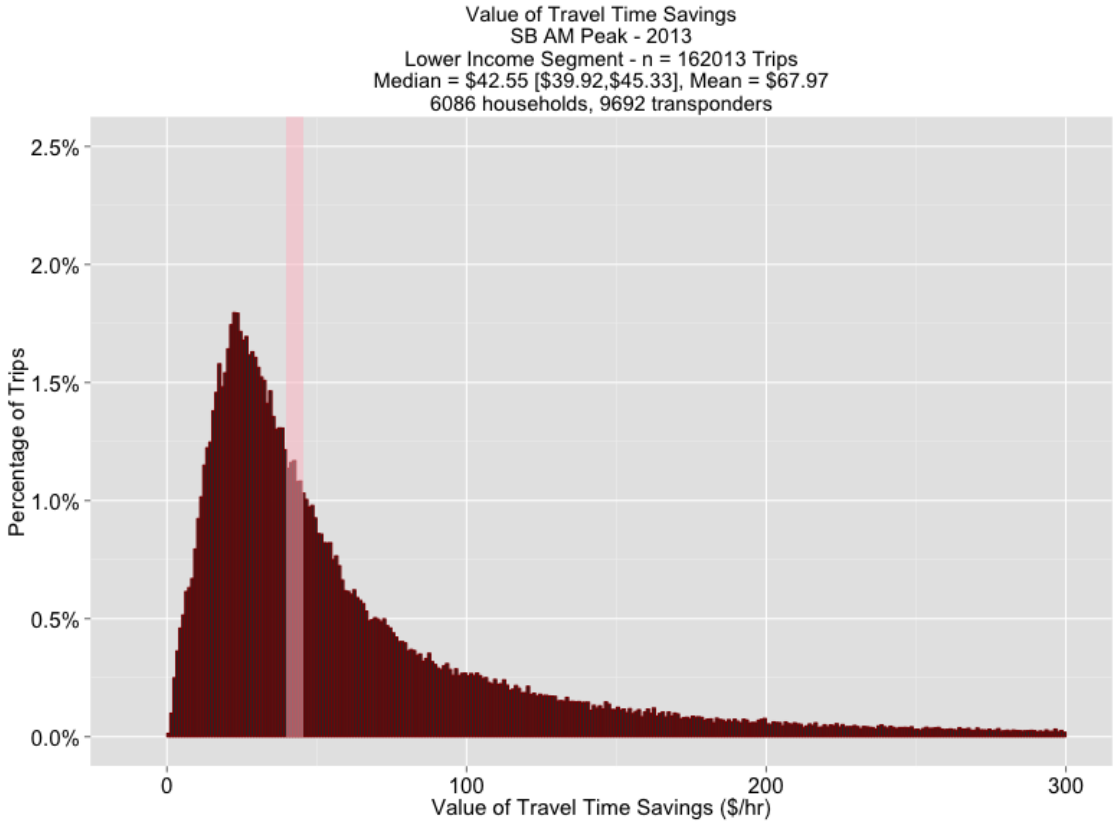


Figure 121: 2013 Southbound VTTS - Lower Income

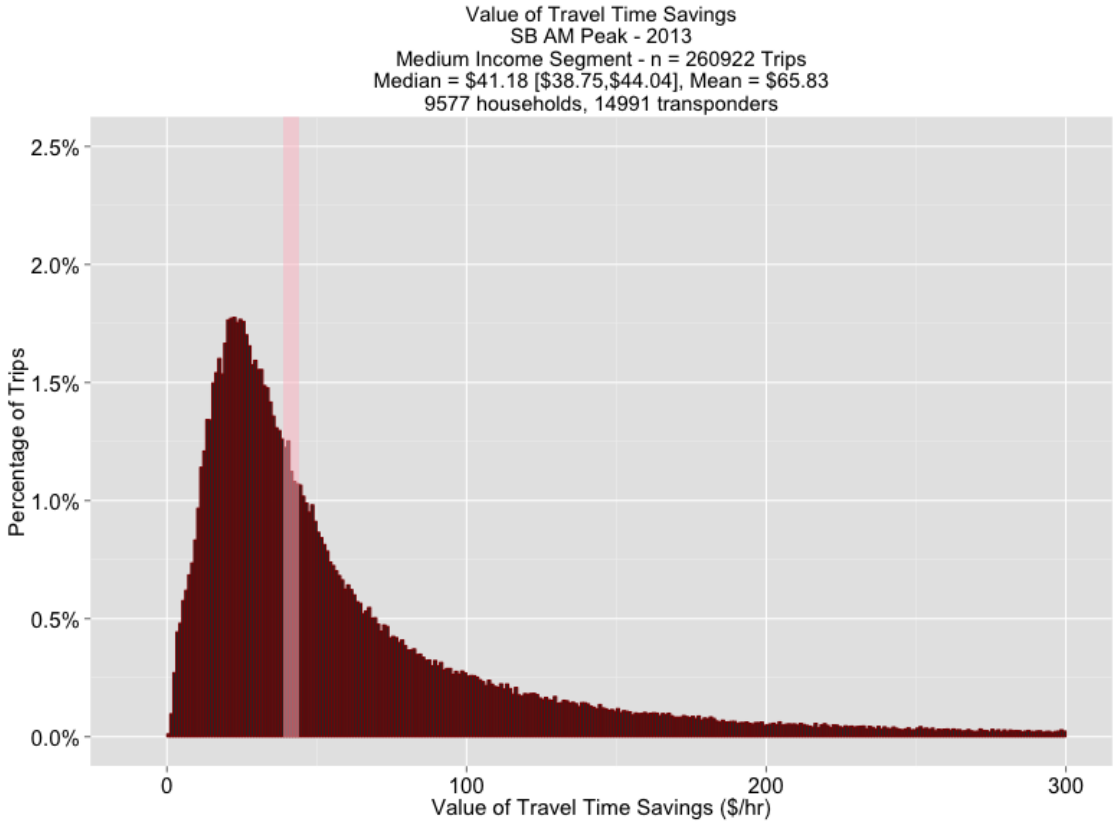


Figure 122: 2013 Southbound VTTS - Medium Income

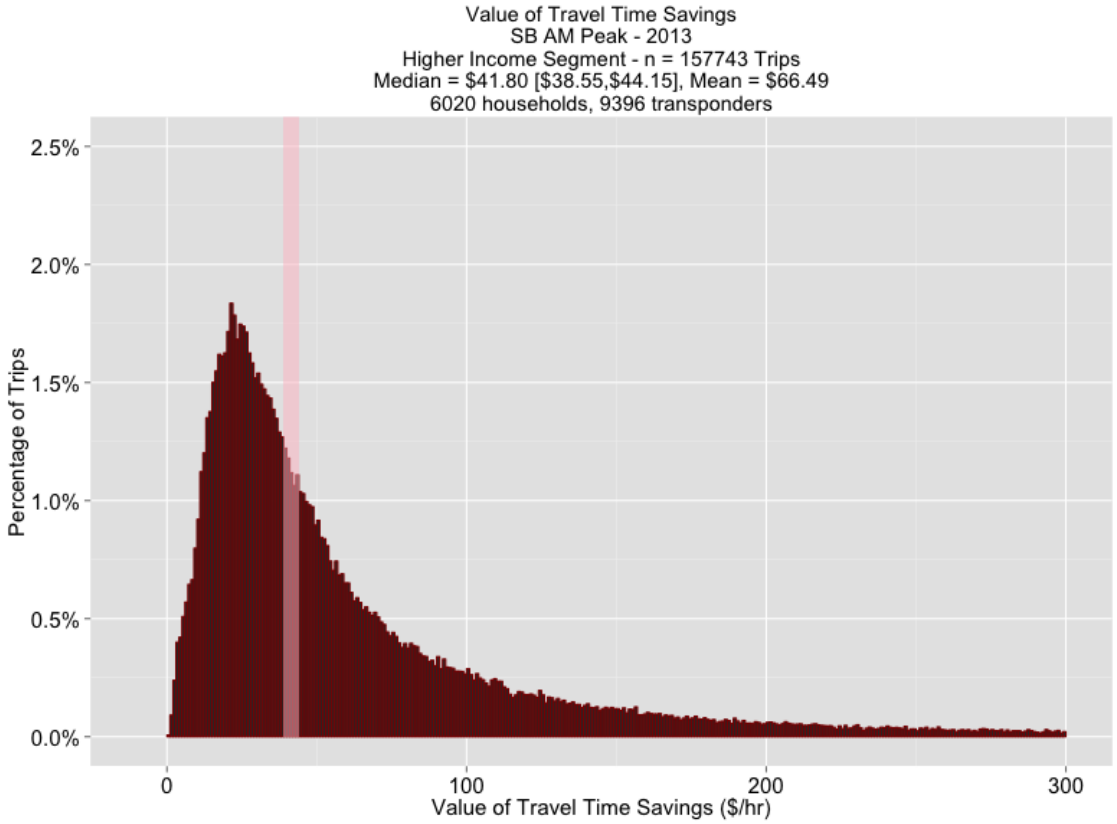


Figure 123: 2013 Southbound VTTS - Higher Income

The next set of figures isolates the differences across the VTTS distributions for the three income groups, starting with a comparison of the lower income segment and the medium income segment. Figure 124 presents these differences at the $1 bin level; in the first chart, the medium income distribution has been subtracted from the lower income distribution. The result shows that the medium income segment generally had higher VTTS levels below the $50/hour cutoff mark, after which the lower income distribution was more frequently greater. An important observation to note is the scale of these differences: they remain well below 0.25% of the total across the entire distribution.

The second chart in Figure 124 shows the results of subtracting the higher income value of travel time savings 2013 distribution from the lower income distribution. Here the pattern differs slightly; the lower income distribution includes more trips below approximately $10/hour and above roughly $45/hour. The higher income segment includes more trips within that interval. Again, the differences are very small in magnitude, remaining below 0.25% for each bin.

The final chart in Figure 124 compares the medium and higher income segment distributions; in this case the higher income distribution was subtracted from the medium income distribution. The largest differences occur below the $50/hour mark, but again the scale of the differences is very small. While the medium income segment includes more trips below roughly $10/hour, the size of the distributional differences makes it difficult to make any meaningful inferences about behavioral differences among the users in the different income segments.

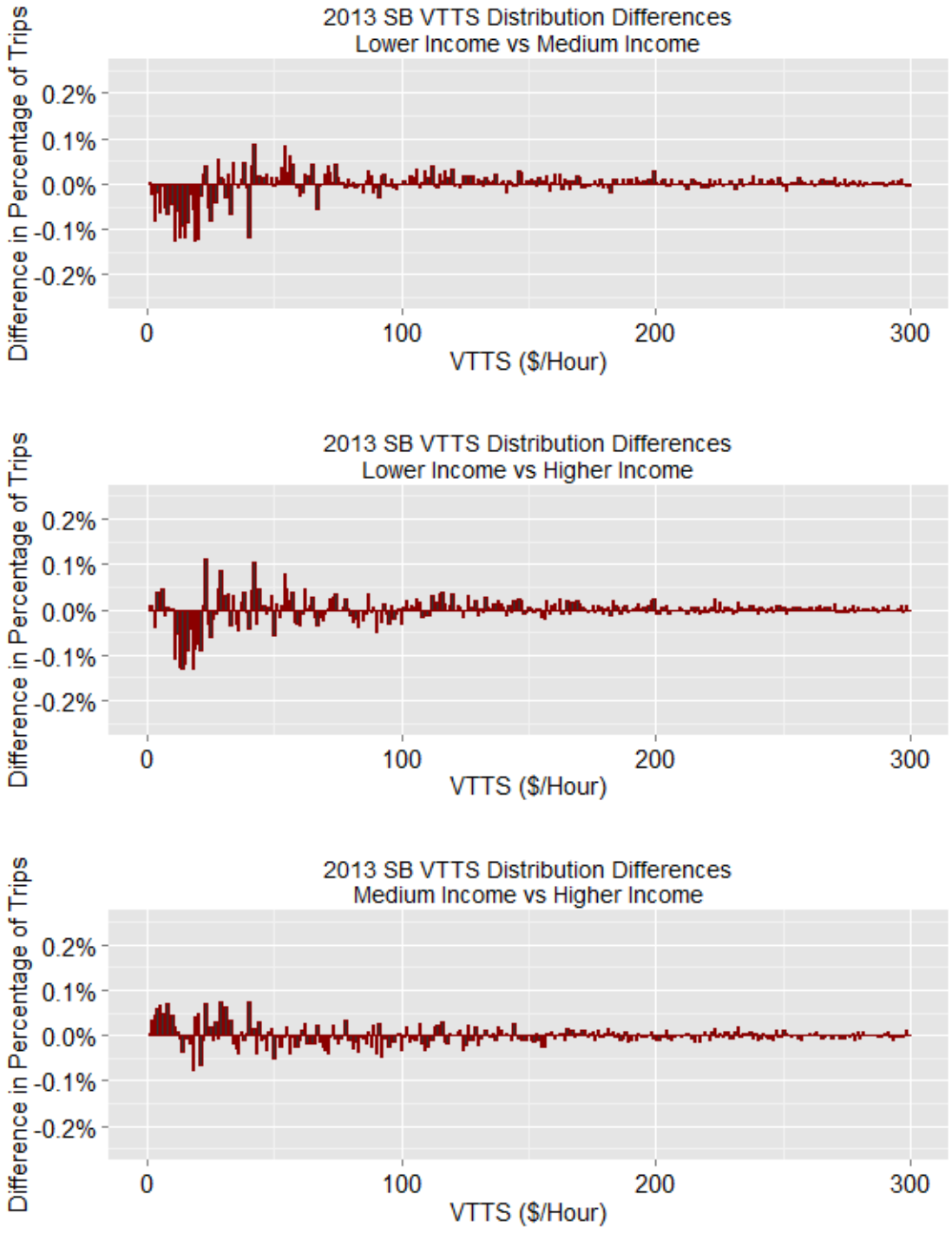


Figure 124: 2013 Southbound VTTS Differences

2013 Northbound Distributions and Differences

The next set of figures consists of the value of travel time savings distributions for northbound PM-peak period trips in 2013. Figure 125 illustrates the distribution for the lower income segment, and represents 181,275 trips by 11,948 transponders and 7,318 households. The median value of the resulting distribution is well below that of the initial analysis, which examined corridor-length trips by all users for nine months across late 2012 and early 2013. That distribution had a median value of $25.66, over $7 higher than the median exhibited here.

Figure 126 presents the medium income distribution, representing 18,179 transponders, 11,398 households, and 291,236 trips. Again, the distribution is more concentrated at the lower end of the value of travel time savings spectrum than Figure 116 above. Unlike the southbound VTTS results, the median and mean values of this middle income segment are higher than those of the lower and higher income segments for the northbound PM-peak trips.

Figure 127 shows the northbound peak period distribution for higher income segment households in 2013. The 169,600 trips were taken by 12,028 unique transponders and 7,630 households. Here the mean and median measures are the lowest of the three income segments, but again the differences are slight. The centrality measures for all three income segments were lower than all of the northbound values in Table 55, including those for the infrequent users, which were the lowest VTTS values in the whole study.

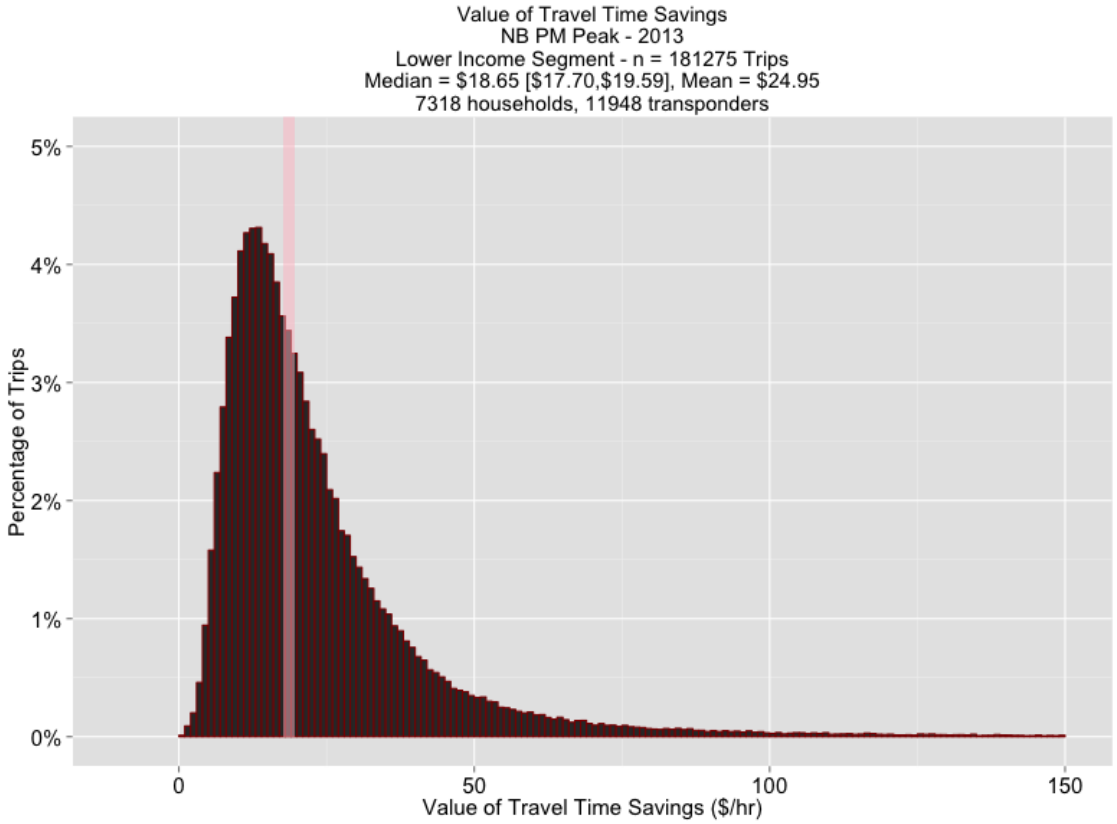


Figure 125: 2013 Northbound VTTS - Lower Income

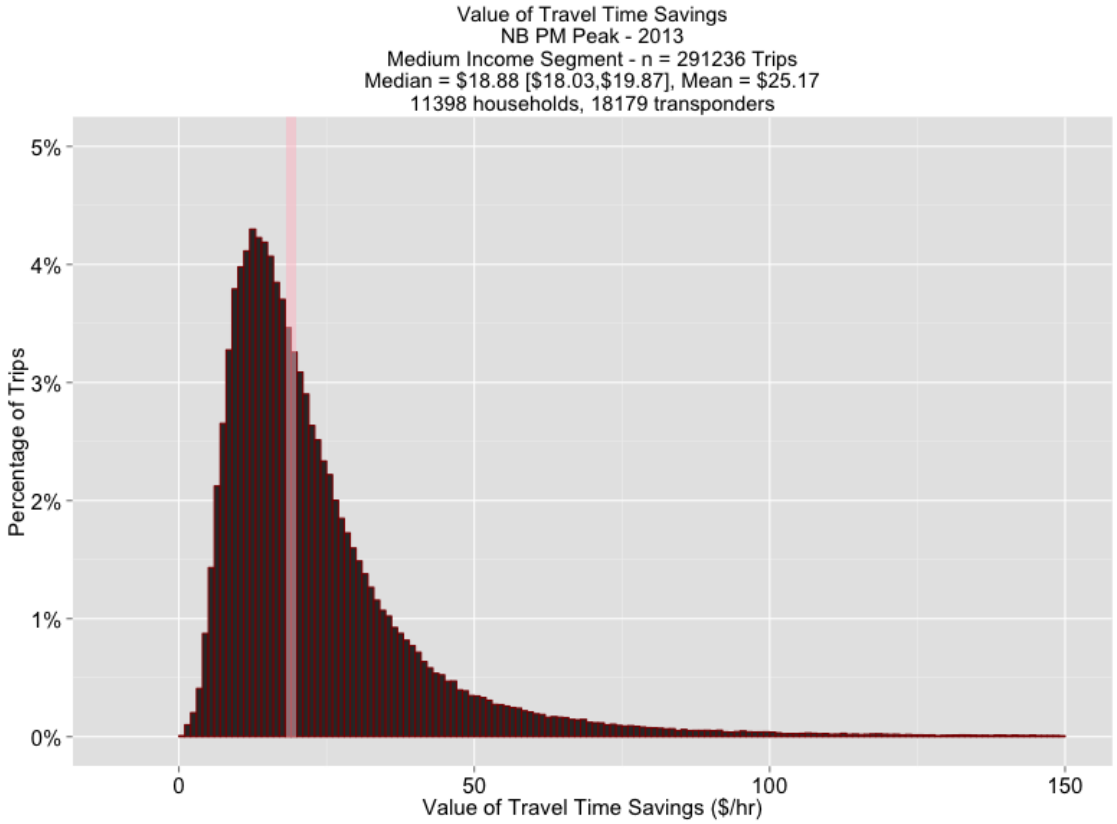


Figure 126: 2013 Northbound VTTS - Medium Income

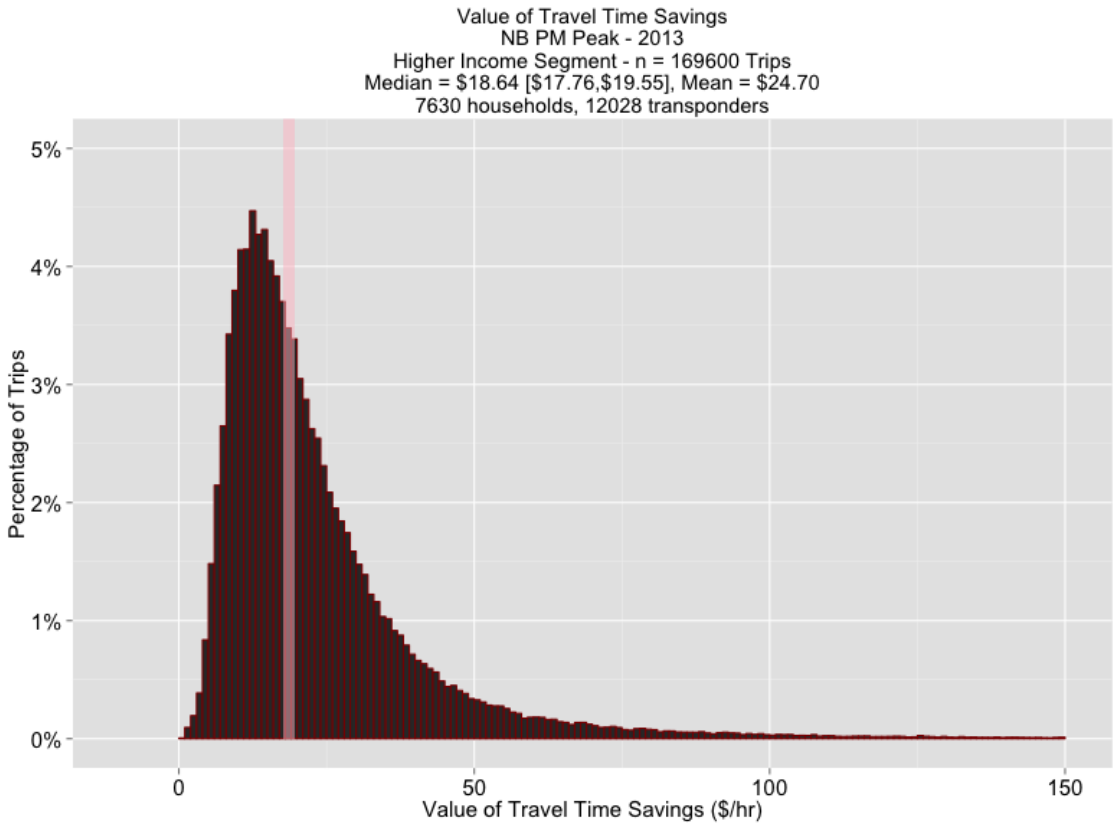


Figure 127: 2013 Northbound VTTS - Higher Income

The next figure presents the differences in the distributions for the northbound peak Express Lane trips. Again the magnitudes of the differences are very small: all are below the 0.2% line on the charts below. The patterns that can be discerned from these plots say very little due to the very minor scale of the discrepancies. The first plot illustrates the results of the medium income segment subtracted from the lower income segment; here the lower income segment appears to have more trips in the VTTS range between $0 and $10/hour. The second chart, in which the higher income VTTS distribution is subtracted from the lower, sees more higher-income trips between roughly $10/hour and $30/hour. In comparing the medium and higher income segments, it appears that the higher income segment has a higher proportion of trips occurring with a

VTTS value below $25/hour. Once again, however, the scale of the differences indicates that even these patterns have little if any practical impact.

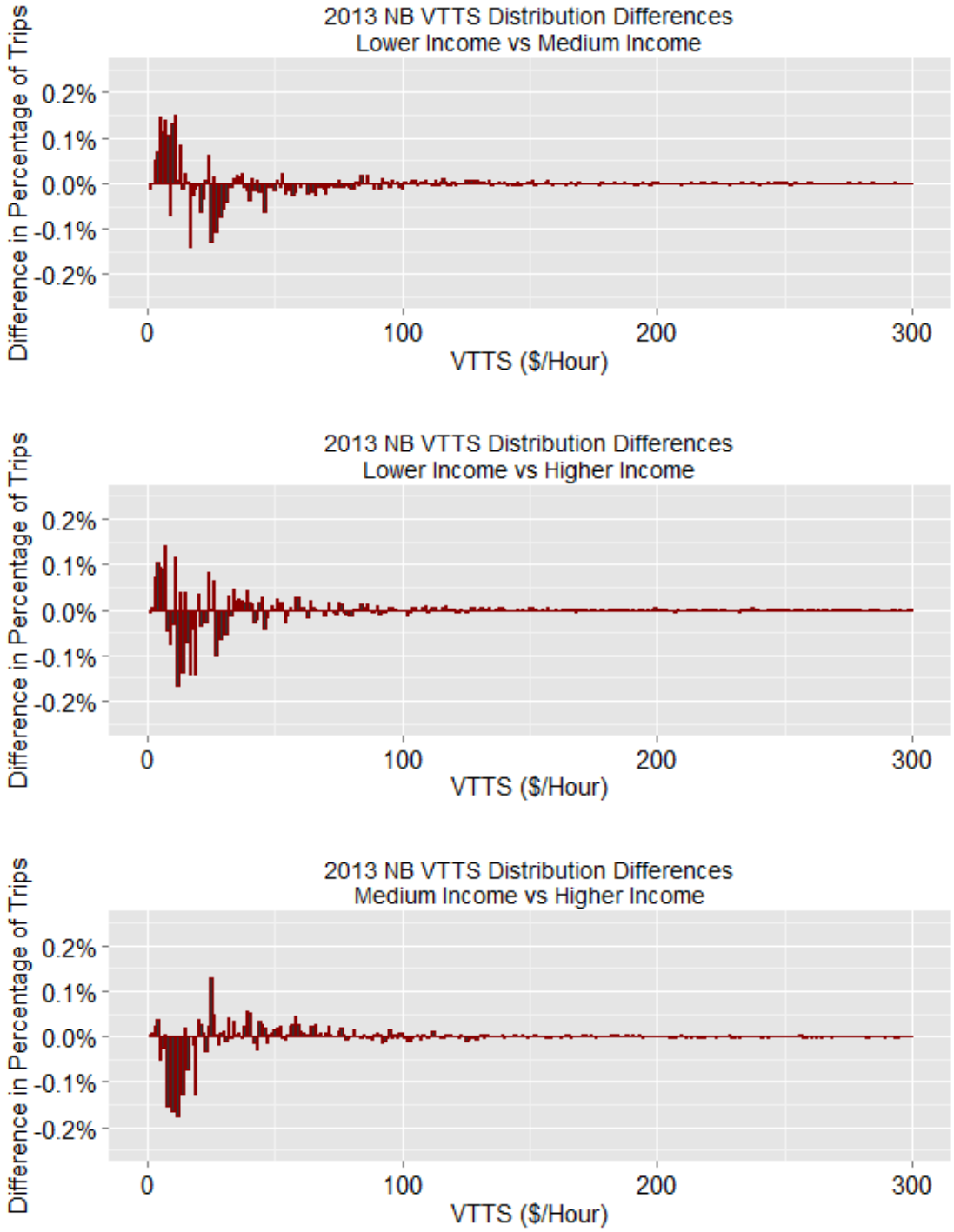


Figure 128: 2013 Northbound VTTS Differences

Overview of Income Segment Differences in Value of Travel Time Savings

The most striking observation regarding the value of travel time savings data for the three income segments presented above is the similarity in their distributions. Table 58 presents an overview of the results for the different income segments, including the inter-quartile range values and the skewness and kurtosis of each segment. The various measures differ only slightly between the three segments in the two different directions. Inter-quartile ranges are much larger in the southbound direction, ranging from $54.06 to $55.94. Similar measures in the northbound direction range from $16.28 to $16.59. The measures of the shapes of the distributions in the form of the skewness values indicate long right tails in all cases, with this lack of symmetry more pronounced in the northbound direction. Similarly, the kurtosis results, while similar across income segments within a direction, indicate much more peaked-ness in the northbound results. The bootstrapped confidence intervals, which represent the 25^{th} and 975^{th} median values calculated from 1000 iterations, overlap in both the southbound and northbound directions. This overlap further illustrates the similarities in the distributions; the medians cannot be said to be different at the 95% confidence level.

To support the visual inspection of the distributions, Mann-Whitney tests were used to investigate whether the various distributions were equal. In most cases, the null hypothesis was rejected at well over the $\alpha=0.05$ confidence level. The only exception was the comparison of the higher and lower income segment distributions in the northbound PM peak; here the null hypothesis could not be rejected. Regardless of the results of the Mann-Whitney tests, the table illustrates, as the previous plots did, that the practical differences in dollars between the distributions are small.

To the extent that these results provide insight into Express Lane user behavior, they suggest that users in different household income groups will pay very similar amounts of money to save a given amount of time. As discussed in the Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter, there are many different sources of bias in the data set. Additionally, the measure itself includes its own set of limitations: users do not know how much time they will save before choosing to use the Express Lanes, for one. The method of VTTS estimation here also does not account for non-linear effects, and so implies that drivers use the same decision-making process in choosing to save one second or five minutes. In light of the previous observation about the lack of foreknowledge regarding travel time savings, this limitation may be justifiable. It is also unlikely that drivers think of the cost of their trip in terms of dollars per hour, especially those trips where the VTTS exceeds $100/hour. All of these considerations must be acknowledged when interpreting these results.

Table 58: Summary Table of 2013 VTTS Distributions by Income Segment

	Southbound			Northbound		
	Lower	Medium	Higher	Lower	Medium	Higher
Number of Trips	162,013	260,922	157,743	181,275	291,236	169,600
Number of Transponders	9,692	14,991	9,396	11,948	18,179	12,028
Number of Households	6,086	9,577	6,020	7,318	11,398	7,630
Median VTTS	$42.55	$41.18	$41.80	$18.65	$18.88	$18.64
25th Percentile	$24.71	$23.94	$24.18	$12.32	$12.52	$12.42
75th Percentile	$80.65	$78.00	$79.03	$28.83	$29.11	$28.70
Bootstrapped Confidence Intervals for Sample Median	[$39.92, $45.53]	[$38.75, $44.04]	[$38.55, $44.15]	[$17.70, $19.59]	[$18.03, $19.87]	[$17.76, $19.55]
Mean VTTS	$67.97	$65.82	$66.49	$24.95	$25.17	$24.70
Skewness	2.61	2.69	2.67	6.43	6.49	6.37
Kurtosis	11.17	11.82	11.61	72.08	73.87	74.09
Mann-Whitney: Versus Lower	N/A	$p<2.2x10^{-16}$	$p=3.28x10^{-10}$	N/A	$p=1.35x10^{-11}$	p=0.579
Mann-Whitney: Versus Medium	$p<2.2x10^{-16}$	N/A	$p=6.45x10^{-6}$	$p=1.35x10^{-11}$	N/A	$p=1.37x10^{-9}$
Mann-Whitney: Versus Higher	$p=3.28x10^{-10}$	$p=6.45x10^{-6}$	N/A	p=0.579	$p=1.37x10^{-9}$	N/A

Full-Length Trips versus Partial Trips

One of the objectives of this expansion of the previous value of travel time savings analysis is to compare the results for users who traverse the entire corridor with the results of those users who only use a portion of the corridor. In particular, researchers were interested in those users who, during southbound trips, leave the HOT facility before the I-285 interchange, and those northbound users who exit the facility after the recurring congestion prior to the Jimmy Carter Boulevard exit. This section of the chapter will focus on those partial trips and compare their results to those of full-corridor users. Corridor trips are divided into four categories: through trips (which traverse the entire duration of the Express Lanes from Old Peachtree Road to I-285 and vice versa), those that begin at one endpoint (Old Peachtree Road for southbound trips, I-285 for Northbound trips) and end before the next endpoint, those that begin within the corridor (not an endpoint) and continue until the end (I-285 for southbound trips, Old Peachtree Road for northbound trips), and finally those that begin and end within the corridor. Again, this chapter presents the results from 2013.

2013 Southbound Full Length versus Partial Trips Comparison

Table 59 shows the number of southbound trips for 2013 that are full-corridor trips, those that start at Old Peachtree Road and end before I-285, those that start after Old Peachtree Road and end at I-285, and those that start after Old Peachtree Road and end before I-285. Because this data set requires both HOT and GP lane data, trips that originate on SR-316 are excluded. Each category of partial trips far outnumbers the full trips by 15,627-142,179 trips; the three partial categories include more transponders, households, and more frequent trips per transponder as well. Trip frequency per transponder is lowest

in the full trip category, though the average speed is greater. The average trip speed is lowest in the category of trips that occur within the corridor without including either the northern-most or southern-most endpoints.

Table 59: Overview of Southbound 2013 Trips by Length

	Full Trips (OPS to 285S)	Partial Trips Start at OPS End Before 285S	Partial Trips Start after OPS, End at 285S	Partial Trips Start and End between OPS and 285S
Number of Trips	71,766	87,393	207,574	213,945
Number of Transponders	10,975	11,484	22,086	19,780
Number of Households	7,898	8,197	15,835	14,320
Average Trips per Transponder	6.54	7.61	9.40	10.82
Average Trip Speed (mph)	55.4	54.2	53.0	49.5

The value of travel time savings distributions for full trips, Old Peachtree to mid-corridor trips, mid-corridor trips to I-285, and mid-corridor to mid-corridor trips are presented in Figure 129 through Figure 132 respectively. Unlike the previous income-segmented distributions, these charts illustrate distinct differences among the four trip lengths. In particular, those trips that begin after Old Peachtree Road see substantially higher mean and median VTTS measures. These two categories, Mid-285 and Mid-Mid, see far more trips at the high end of the VTTS spectrum. The two categories of trips that begin at Old Peachtree Road, those that end at I-285 and those that end mid-corridor, more closely resemble each other than the remaining two trip categories.

Once again, the author performed Mann-Whitney tests to compare the distributions. In each case, the null hypothesis of distributional equality was rejected at well over the $\alpha = 0.01$ confidence level. As before, this is likely an effect of the size of each sample for the four different trip categories. Unlike the income-segmented trips, the

practical differences between the distributions are more apparent among these four categories. The mid-corridor to mid-corridor trips yield the highest median VTTS results, and the mean VTTS value for that category is $40.85 higher than that of the lowest category (Old Peachtree to mid-corridor) and $11.17 higher than that of the second-highest category (mid-corridor to I-285). The bootstrapped confidence intervals overlap between the full length and Old Peachtree to mid-corridor categories, as well as between the mid-corridor to I-285 and mid-corridor to mid-corridor categories. The prevailing theme among these four distributions is that those trips that begin at similar points on the corridor, either at Old Peachtree or elsewhere within the corridor, are similar, regardless of the end points.

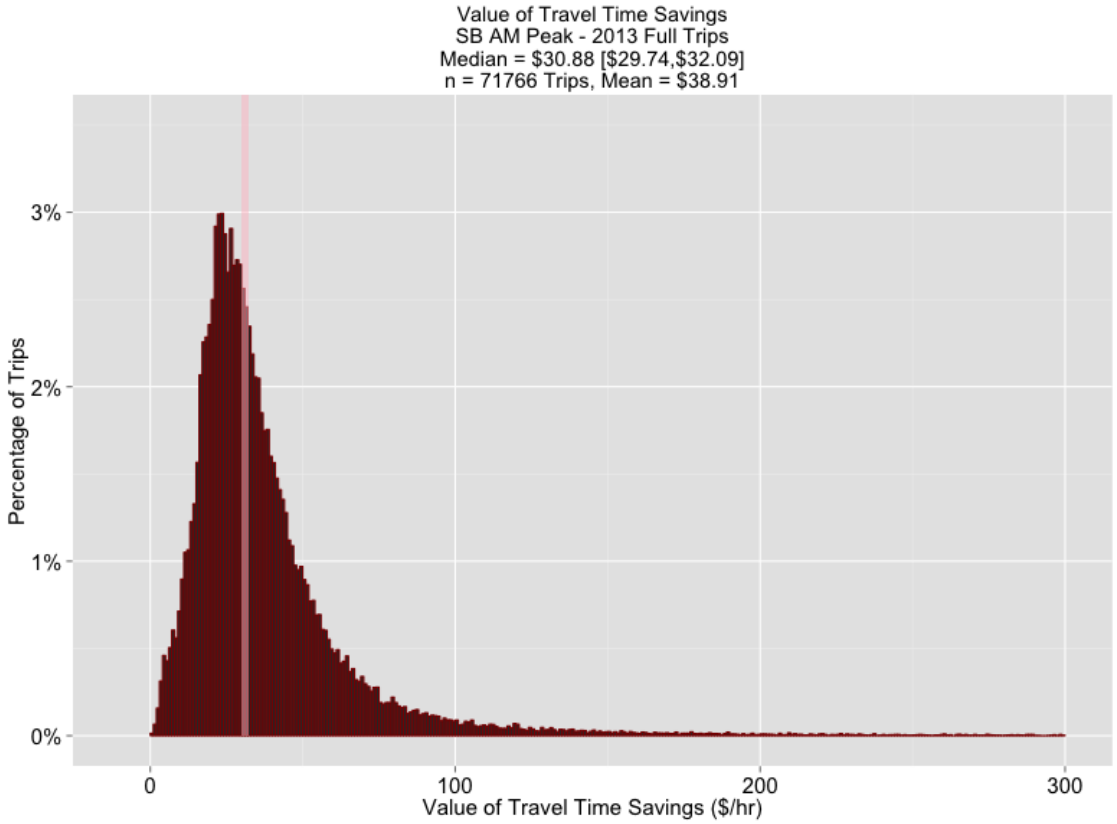


Figure 129: 2013 Southbound VTTS - Full Length Trips

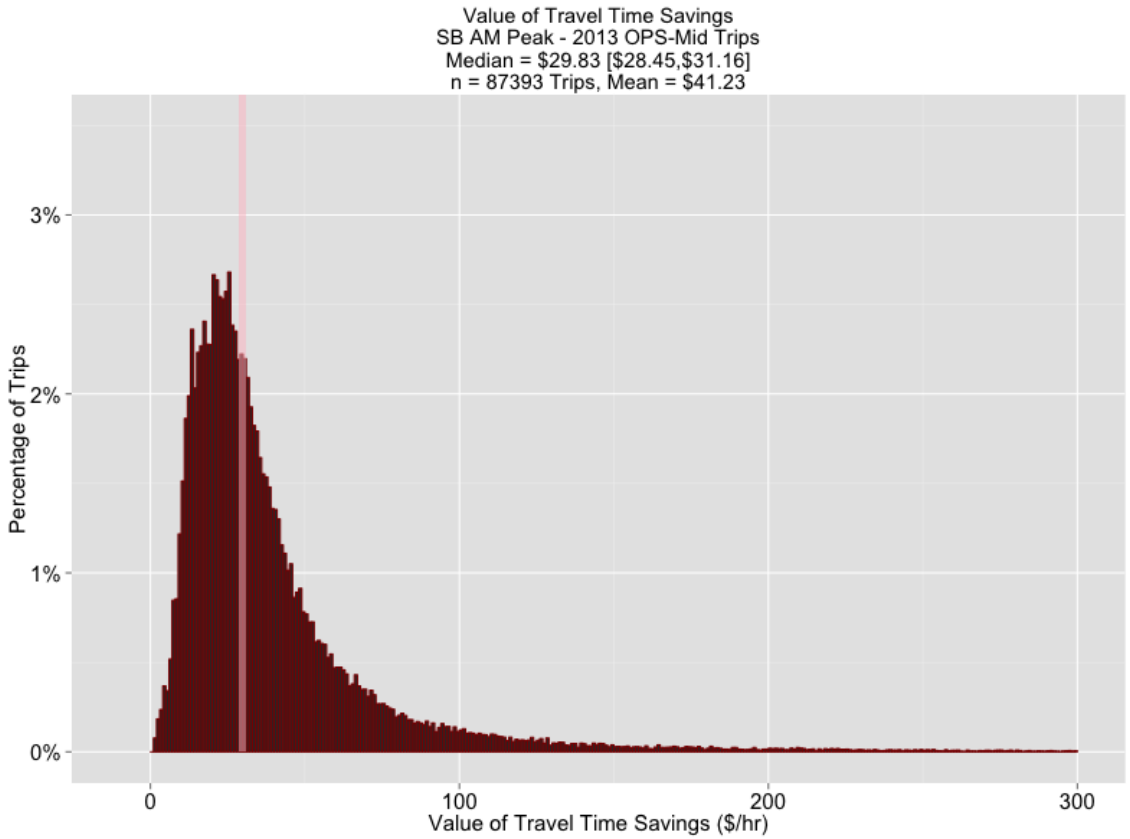


Figure 130: 2013 Southbound VTTS – Old Peachtree to Mid-Corridor Trips

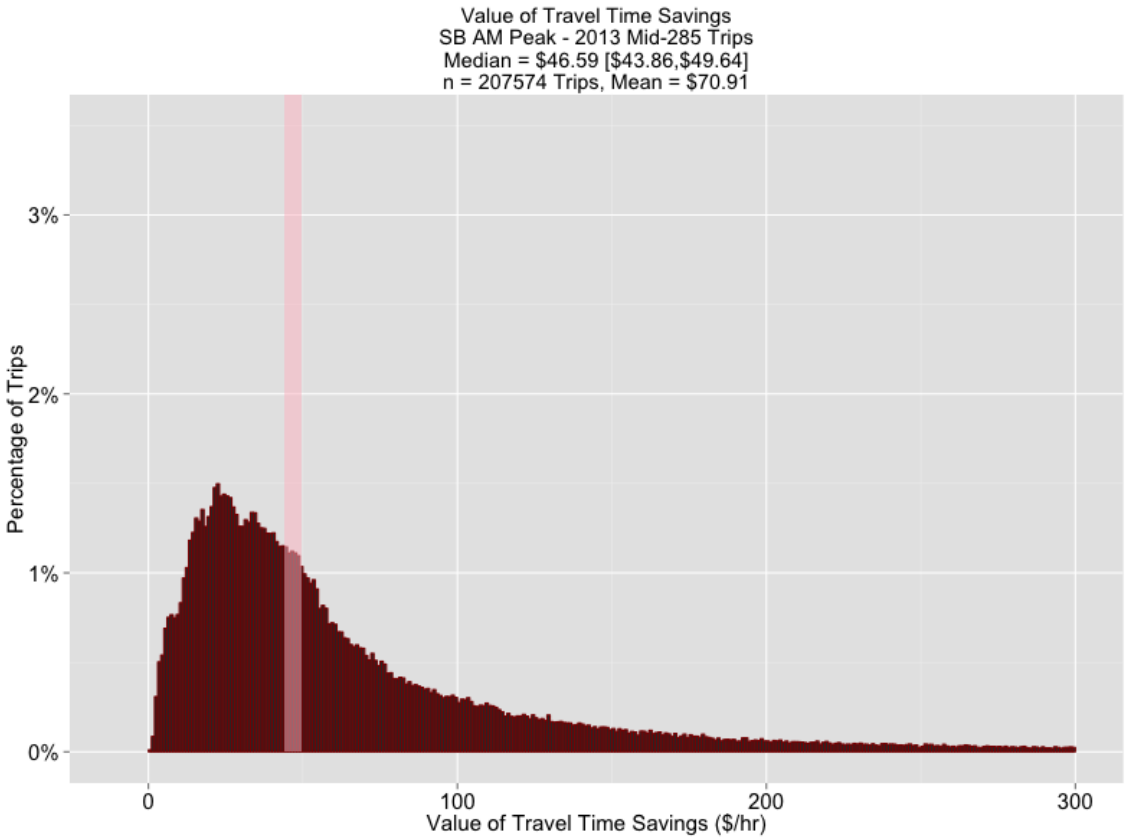


Figure 131: 2013 Southbound VTTS – Mid-Corridor to I-285 Trips

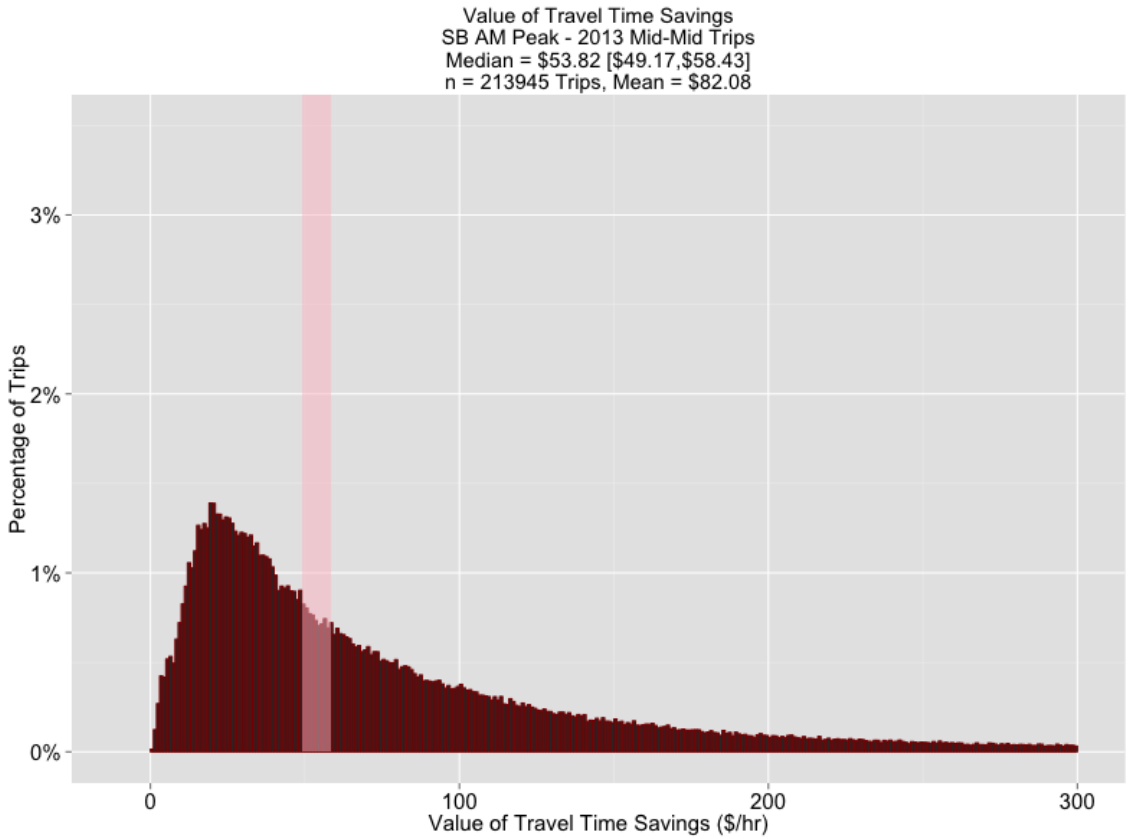


Figure 132: 2013 Southbound VTTS - Mid-Corridor to Mid-Corridor Trips

2013 Northbound Full Length versus Partial Trip Comparison

Table 60 presents an overview of northbound trips for 2013 that are full-corridor trips, those that start at I-285 northbound and end before Old Peachtree Road, those that start mid-corridor and end at Old Peachtree Road, and those that start mid-corridor and end mid-corridor. Unlike the southbound trips, the corridor-length trips in the northbound direction are not the least numerous of the four categories. The least frequent trips are those that start mid-corridor and end at Old Peachtree Road. As was shown in Chapter 3, the majority of northbound trips begin at the I-285 segment: the start of the corridor. While the mid-corridor to Old Peachtree trips are the least common, they do yield the highest average trip speeds. The most frequently observed category of trips, those that

start at I-285 and end mid-corridor, have the highest per-transponder trip frequency and the lowest average trip speed.

Table 60: Overview of Northbound 2013 Trips by Length

	Full Trips (285N-OPN)	Partial Trips Start at 285N, End Before OPN	Partial Trips Start after 285N, End at OPN	Partial Trips Start and End between 285N and OPN
Number of Trips	111,027	294,938	97,540	138,606
Number of Transponders	15,852	26,090	15,147	23,020
Number of Households	11,259	18,230	10,923	16,911
Trips per Transponder	7.00	11.30	6.44	6.02
Average Trip Speed (mph)	60.7	55.0	61.9	55.2

Figure 133 through Figure 136 present the VTTS distributions for the four categories of northbound 2013 trips outlined above. The VTTS values in these four distributions are lower overall than their southbound counterparts; the highest mean northbound VTTS, from mid-corridor to Old Peachtree Road, is only $1.27 higher than the lowest southbound mean VTTS (that of the full length trip category). Those users who begin at I-285 and exit the Express Lanes before Old Peachtree road exhibit the lowest mean and median VTTS values. The bootstrapped confidence intervals for the sample medians are less similar in the northbound direction. Whereas the southbound trips had two pairs of overlapping confidence intervals, in the northbound direction only one pair of confidence intervals overlap: those of the full length and mid-corridor to mid-corridor trips. As was the case in the southbound direction, Mann-Whitney distributional comparison tests reject the null hypothesis of distributional equality among all four trip categories. Again, with a minimum of 97,540 observations in each category, this was

expected. Still, the practical differences between the distributions are more stark in the northbound direction.

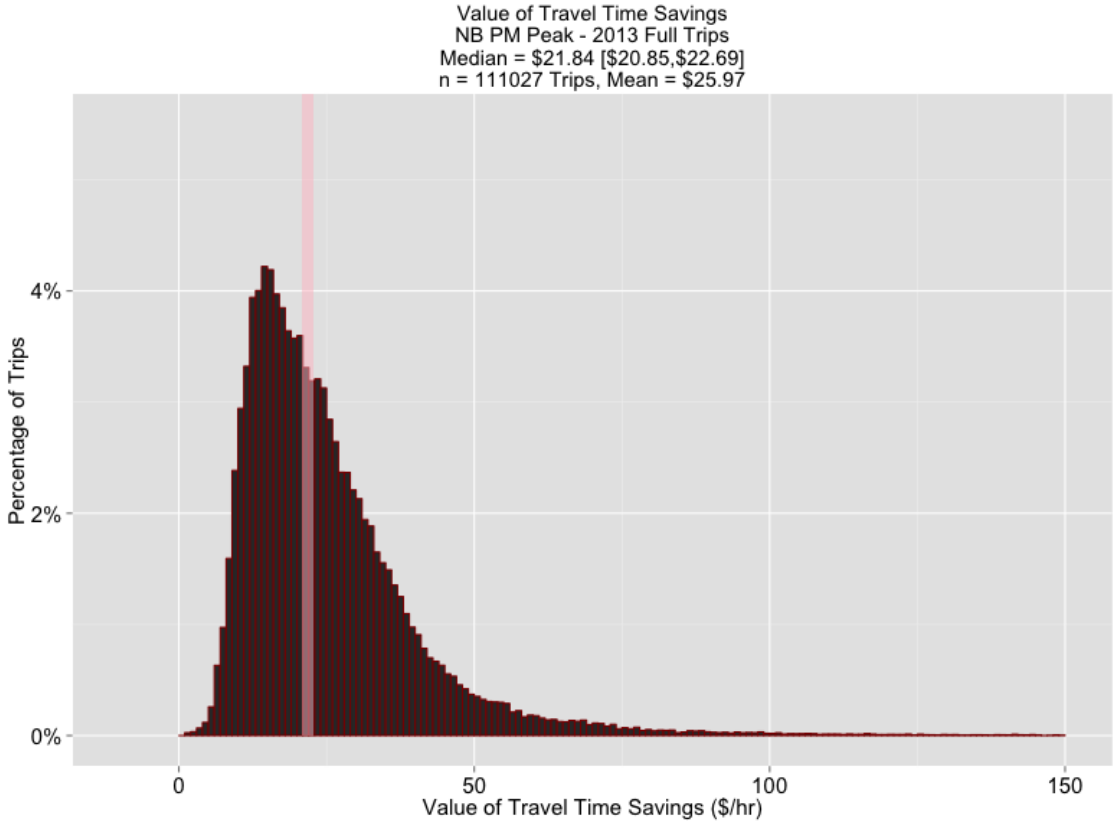


Figure 133: 2013 Northbound VTTS - Full Length Trips

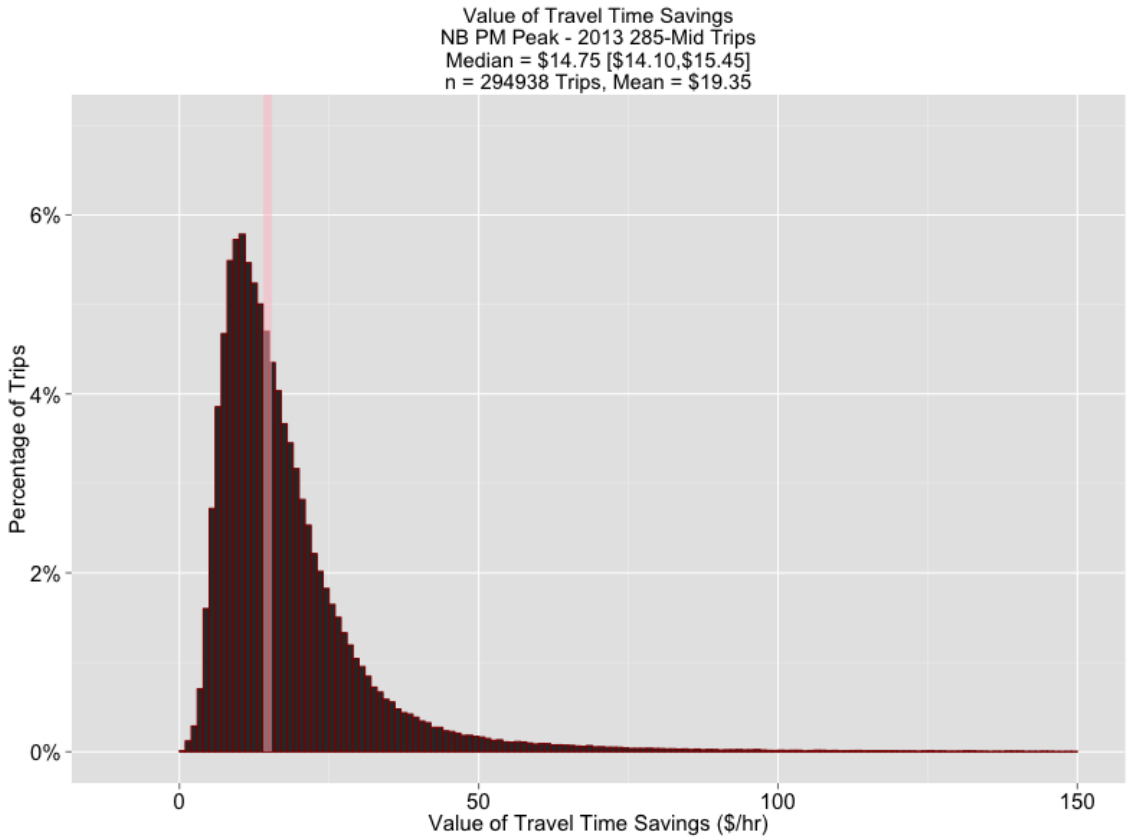


Figure 134: 2013 Northbound VTTS – I-285 to Mid-Corridor Trips

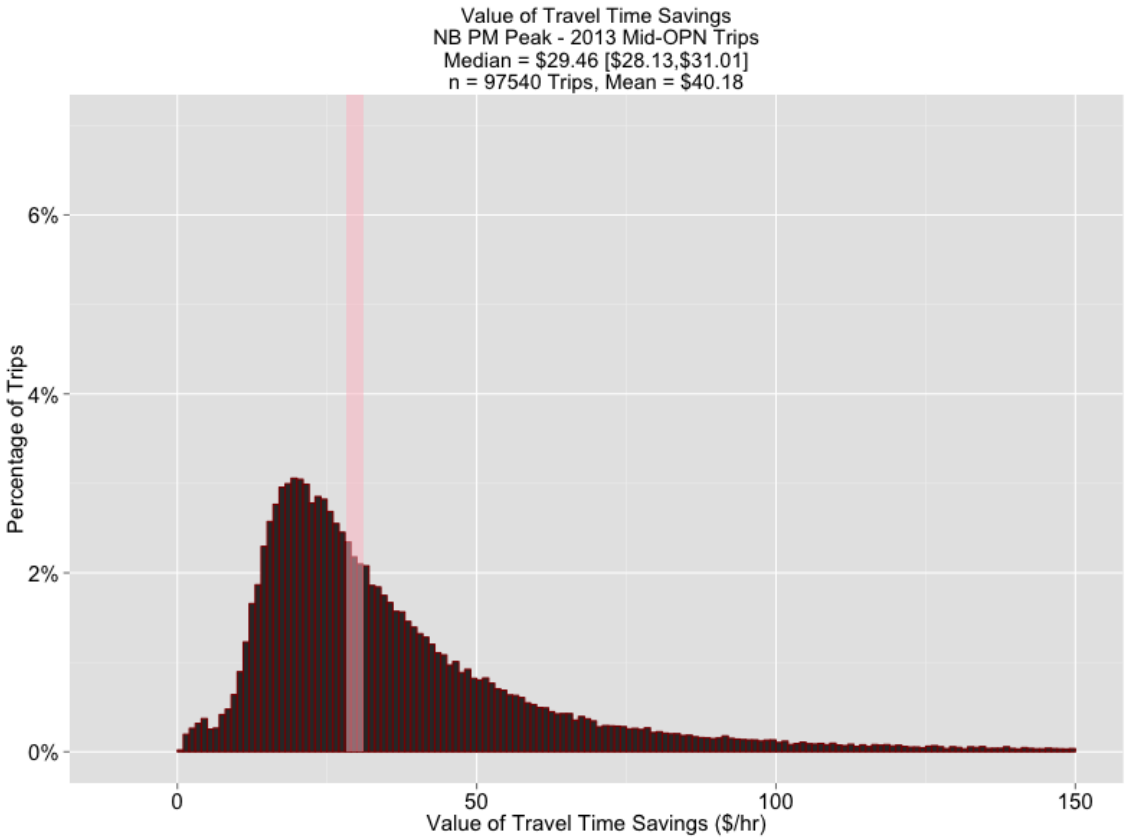


Figure 135: 2013 Northbound VTTS – Mid-Corridor to Old Peachtree Road Trips

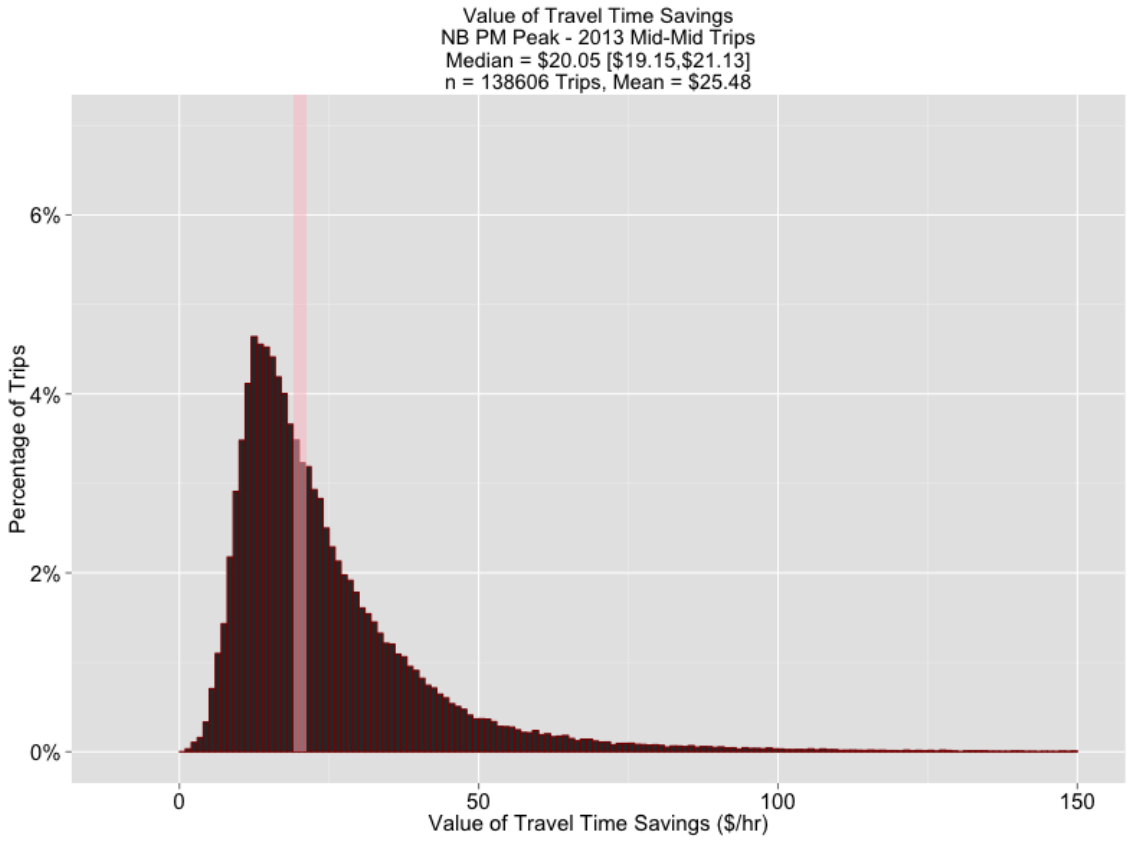


Figure 136: 2013 Northbound VTTS - Mid-Corridor to Mid-Corridor Trips

Full-Length versus Partial Trip Comparison Summary

The differences among the value of travel time savings distributions for different trip lengths are far more pronounced than those among different income segments. Table 61 provides an overview of the distributional measures for the different trip lengths and directions. Whereas the previous section revealed little variation between low, medium, and high income users of the toll lanes, this section illustrated distinct differences between full length trips and partial length trips. As in the income-based distributions, the southbound and northbound behaviors were different as well. The southbound data revealed that users taking partial length trips pay more to save less time than those who stay in the corridor for the duration. Users who began their trips at similar locations, either at Old Peachtree Road or mid-corridor, exhibited similar VTTS distributions. The bootstrapped median confidence intervals supported this, as they overlapped for the full length and Old Peachtree to mid-corridor trips, and also for the mid-corridor to 285 and mid-corridor to mid-corridor trip.

In the northbound direction, these relationships are different. Users who enter the Express Lanes mid-corridor and exit at Old Peachtree Road more frequently pay higher tolls to save less time, as evidenced by the larger tail of the full-trip VTTS distribution. In this case, it is the users who begin and I-285 and exit the lane before Old Peachtree, short of the full corridor length, who have the lowest mean and median VTTS values. Users often do not need to continue in the HOT lane location northward of the Jimmy Carter Boulevard segment, the second in the sequence, as the speed differences between the HOT and GP lanes decline significantly once congestion associated with the I-85/I-285 merge is complete. The distributions are all narrower and more heavily weighted at

the lower end than the southbound figures, indicating more time saved and/or lower tails paid in the northbound direction. The differences within the four categories are greater as well; only two (full length trips and mid-corridor to mid-corridor trips) exhibit overlapping confidence intervals around the median.

These comparisons speak to differences in the characteristics and behavior of northbound and southbound toll lane trips and trip-takers. Southbound trips have the highest 'value' when users stay in the lane for the duration; that is, those trips deliver more travel time savings for less cost. This is not the case for the northbound trips, where the full length trips provide higher tolls and lower time savings than those trips that end before Old Peachtree Road. In the northbound direction, the most 'value' can be found in entering the toll lane at its I-285 beginning and exiting prior to the end of the corridor. This is likely due to differing congestion patterns and driver behavior in the morning and afternoon peak periods, as well as different toll schedules for the different directions and times.

Table 61: Summary Table of 2013 VTTS Distributions by Trip Length

	Southbound				Northbound			
	Full Length	OPS-Mid	Mid-285	Mid-Mid	Full Length	285-Mid	Mid-OPN	Mid-Mid
Number of Trips	71,766	87,393	207,574	213,945	111,027	294,938	97,540	138,606
Number of Transponders	10,975	11,484	22,086	19,780	15,852	26,090	15,147	23,020
Number of Households	7,898	8,197	15,835	14,320	11,259	18,230	10,923	16,911
Median VTTS	$30.88	$29.83	$46.59	$53.82	$21.84	$14.75	29.46	20.05
25th Percentile	$21.92	$19.69	$26.05	$27.89	$15.13	$9.99	20.01	13.86
75th Percentile	$44.70	$46.43	$85.48	$105.26	$31.23	$21.98	46.14	30.32
Bootstrapped Confidence Intervals for Sample Median	[$29.74, $32.09]	[$29.83, $31.16]	[$43.86, $49.64]	[$49.17, $58.43]	[$20.85, $22.69]	[$14.10, $15.45]	[$28.13, $31.01]	[$19.15, $21.13]
Mean VTTS	$38.91	$41.23	$70.91	$82.08	$25.97	$19.35	$40.18	$25.48
Skewness	4.95	4.43	2.56	2.12	6.42	8.51	4.59	6.03
Kurtosis	41.77	31.41	10.90	8.13	89.61	121.19	34.99	74.71
Mann-Whitney: Versus Full	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Mann-Whitney: Versus OP-Mid	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Mann-Whitney: Versus Mid-285	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$
Versus Mid-Mid	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	N/A

Chapter Summary

This investigation of the value of travel time savings for I-85 Express Lane users revealed a number of striking findings. Foremost among these was the similarity of the results across the income segments defined in this dissertation. The differences in VTTS results between lower, medium, and higher income households, to the extent that they exist, are marginal at best. These differences most frequently appear on the order of cents rather than dollars. Likewise, it is not the case that higher income households exhibit the highest VTTS results. In the southbound AM-peak trips, the lower income households actually have the highest mean and median values of travel time savings. Overlapping confidence intervals around the median VTTS values indicated that the median values could not be said to be different at the 95% confidence level. It may be the case that any variation in behavior occurs not among these three segments but within a subset of very-high income households, who earn more than the $100,000/year criteria defined for higher income households. This is explored in other chapters of this dissertation, particularly Chapter 12.

The trip length investigation revealed more distinct differences between users who traverse the entire duration of the corridor and those that take partial trips; in that case, the southbound and northbound differences were also more pronounced. In the southbound direction, full-length corridor trips provided higher levels of travel time savings for lower toll rates. In the northbound direction, the toll lanes were most beneficial for those users who entered the Express Lanes at the beginning and left before the end of the corridor. The differences between the distributions of the northbound and southbound VTTS values likely reflect different congestion patterns and trip purposes. In particular, morning trips likely have a

higher proportion of commute trip purposes. Afternoon trips encounter higher levels of congestion near the I-85/I-285 merge, after which the benefits of the Express Lanes decrease.

A major limiting factor in extending the value of travel time savings analysis with the available demographic data is the reduction in sample size that results from the pairing process. As discussed in other chapters of this dissertation, the transponders that can be successfully paired with the Epsilon marketing data make up approximately 20% of the full SRTA transponder population. In addition, this sample is not wholly representative of the population: it is biased towards more frequent users as well as higher income households, households in single family dwelling units, and accounts with one vehicle and one transponder.

The limitations that affect this dataset as a whole extend to this analysis as well. They include the lack of stated preference data and trip data beyond the I-85 corridor. Without a stated preference component to this work, researchers cannot isolate the willingness of users to pay for travel time reliability versus speed, for example. Further, user trip purpose remains unknown for all of the trips studied here. The restriction of trip data to the I-85 corridor paints an incomplete portrait of the trips as well. Users may be willing to pay different toll levels based on the total length of their trips; a user for whom I-85 is just a small portion of a commute may be less concerned with saving time on it. Finally, this analysis is limited to toll lane users only by the nature of its methodology. Elsewhere in this dissertation, the analyses will incorporate choice modeling methods to include unpriced general purpose lane trips in their datasets, thus giving a more complete picture of what corridor users will and will not pay to save travel time.

CHAPTER 11

REGRESSION TREE ANALYSIS

Regression tree analysis can help identify factors that have potential impact on the dependent variable in a model; in this case, that variable is the lane choice decision. The dataset used in this dissertation contains over one hundred variables that may potentially help explain whether a user purchases a toll lane trip or remains in the GP lanes. Regression tree models are used here to narrow down that list. As the dependent variable is a set of discrete values, the analysis used here is strictly known as classification tree rather than regression tree. Both of these methods fall under the larger umbrella of Hierarchical Tree-Based Regression; the interpretation of the results is similar. Regression tree analysis is useful in handling discrete variables with more than two values, and this method also better handles missing data relative to ordinary least squares regression. The method also is unaffected by multicollinearity. Downsides of the method include the inability to identify all correlations, and also the selection of variables that are not causal (Washington, et al., 1997).

This chapter begins with a discussion of the data used in the initial regression tree analysis. The next section discusses the results of the initial analysis and the problematic variables identified. The chapter continues with additional regression tree results without the invalid variables, followed by more variable exploration using the random forest method.

Regression Tree and Random Forest Data

The regression tree analysis presented here uses the constructed trip data from 2013 that were paired with the Epsilon demographic data as discussed in Chapter 5. Table 62 provides an overview of the data set. As shown in the Potential Sample Bias in Paired Vehicle Activity and

Marketing Data chapter, the rate of GP-Only trips decreases as the Epsilon marketing and corridor condition data sets are joined to the base constructed trips. Afternoon trips are more prevalent than AM peak period trips; they also include more Peach Pass transponders from more households.

Table 62: Regression Tree Dataset Overview

	All Trips	AM Southbound Trips	PM Northbound Trips
Total trip count	2,376,450	1,179,060	1,197,390
GP-Only Trips (%)	1,109,409 (46.7%)	559,868 (47.5%)	549,541 (45.9%)
HOT-Only Trips (%)	374,543 (15.8%)	205,987 (17.5%)	168,556 (14.1%)
Mixed Trips (%)	892,498 (37.5%)	413,205 (35.0%)	479,293 (40.0%)
Unique Households	35,073	27,817	33,530
Unique Transponders	63,451	47,253	58,204

Table 63 through Table 67 present the variables that were included in the initial regression tree models. In all of the hierarchical tree-based regression models, the dependent variable is HOT lane use. This variable has a value of zero if the trip never enters the Express Lanes and a value of one if the Express Lanes are used at all during the trip. This includes the full set of trip, operational, and demographic characteristics, as well as certain interaction terms. In all of the regression trees presented here, the complexity parameter was set to 0.01. This means that any variable split must increase the model R^2 goodness of fit measure by 0.01 to be included in the model. In each of the five tables presented, the left-hand column provides the name of the element while the right-hand column lists the variable name used in the regression tree and random forest figures.

Table 63 outlines the trip characteristic variables. Many of these, such as the day of week indicators, are self-explanatory. The toll at daily maximum indicator has a value of one if the toll amount paid by the user is equal to the maximum toll of that day. The HOT toll amount is

the rate paid by the user or what the user would have paid had they chosen the Express Lanes. The corridor segment indicators have a value of one if the user was detected in that corridor segment during their trip, in either the HOT or GP lanes. The segment count is the sum of all of those indicators, and represents the total number of segments over which the trip occurred. Trip distance is calculated from the station numbers of the RFID detection gantries. The hour of day and half-hour of day dummy variables indicate the time at which the trip began. Similarly, the month of year and season indicators represent the month and season in which the trip was taken. The square of toll amount variable simply squares the HOT toll rate, while the maximum daily toll amount represents the highest toll rate recorded for that direction on that day. The trip exit segment variables are dummy indicators that represent the last segment of the user's trip. Similarly, the trip segment path variables present the various combinations of corridor segments. The time since January variable counts the number of months since January of 2013; this is meant to capture the potential effects of deteriorating lane conditions over the course of the year. The dummy variable indicating direction has a value of one when the trip is in the southbound direction; this means the trip occurred during the morning peak period as well (a value of zero mean the northbound trip occurred in the afternoon peak period).

Table 63: Trip Characteristic Variables Included in 2013 Regression Tree Analysis

Trip Characteristics	
Day of week dummy variables	monday, tuesday, wednesday, thursday, friday
Toll at daily maximum dummy variable	tollAtMax
HOT toll amount	tollAmount.HOT
Corridor segment dummy variables	segmentOP, segmentPH, segmentIT, segmentJC, segment285
Count of segments traversed (in either the HOT or GP lanes)	segmentCount
Trip distance (miles)	distancemi
Hour of day dummy variables	sixAm, sevenAm, eightAm, nineAm, threePm, fourPm, fivePm, sixPm
Half hour increment dummy variables	am600, am630, am700, am730, am800, am830, am900, am930, pm1500, pm1530, pm1600, pm1630, pm1700, pm1730, pm1800, pm1830
Month of year dummy variables	january, february, march, april, may, june, july, august, september, october, november, december
Season of year dummy variables	winter, spring, summer, fall
Square of toll amount	tollSquared
Maximum daily toll amount	maxToll
Trip exit segment	end285, endJC, endIT, endPH, endOP
Trip segment path	seg1to1, seg1to2, seg1to3, seg1to4, seg1to5, seg2to2, seg2to3, seg2to4, seg2to5, seg3to3, seg3to4, seg3to5, seg4to4, seg4to5, seg5to5
Time since january	timeSinceJanuary
Direction	southbound

Table 64 presents the corridor condition variables used in the regression tree analysis. The first set of variables, the congested condition dummies, have a value of one when the average general purpose lane speed is below that level (50 mph down to 5 mph). The average speed difference between lane types is the difference in average speeds between the HOT and GP lanes. When a trip occurs across both lane types, that value is the difference over the length of the HOT corridor that the user traverses. The next variable, average speed difference in GP portion of mixed trips, presents the difference in average speeds along the portion of the mixed trip that occurs in the general purpose lanes. The htDensity variable counts the Peach Pass transponders detected in the HOT lane along the length of the user's trip and divides that count

by the distance in miles. The transponder counts for the HOT and GP lanes provide only those counts without controlling for distance. Finally, the square of the average speed difference simply squares the average speed difference between lane types.

Table 64: Corridor Condition Variables

Corridor Conditions	
Congested conditions dummy variables (50mph to 5mph)	congested50, congested45, congested40, congested35, congested30, congested25, congested20, congested15, congested10, congested05
Average speed difference between lane types	htAvgSpeedDiff
Average speed difference in GP portion of mixed trips	gpAvgSpeedDiff
htDensity (vehicles per mile in HOT lane over 15 minutes)	htDensity
Square of average speed difference	avgSpeedDiffSquared
Transponder Counts	htTransponderCount, gpTransponderCount

The household characteristic variables are listed in Table 65. The first, race/ethnicity, indicates the racial makeup of the household. The next sets of variables, indicating the presence of adults and children of various ages, report the existence of individuals within the given age ranges. Similarly, presence of one child and presence of multiple children dummy indicators are based off of the number of children variable. The physical structure of the household is represented by the living area square footage, property lot size in acres, age of home, and dwelling type variables. Household income is represented three ways: the three-group segmentation, the five group segmentation, and as an ordinal variable. Household education lists the average level of education completed by the adults in the household. Home ownership status indicates whether the individuals are likely or probably renters or owners. Marital status indicates whether the occupants are married, while household age represents the age of the head of the household. Length of residence indicates the amount of time individuals with a given last name have been present in the house. Occupation is a categorical variable with twenty-one

different values. Household size presents the total number of individuals, while family composition provides sixteen different potential values of family types. The trip count measure was included to examine whether frequent corridor users may behave differently than infrequent users. This measure counts both HOT and GP trips. Note that this measure occurs at the transponder level rather than the household level.

Table 65: Household Characteristic Variables

Household Characteristics	
Race/Ethnicity	raceWhite, raceBlack, raceHispanic, raceAsian, raceOther
Presence of adults – unknown age	presence.of.adults.unknown.age.enhanced
Presence of adults – 75+	presence.of.adults.age.75.specific.enhanced
Presence of adults – 65-74	presence.of.adults.age.65.74.specific.enhanced
Presence of adults – 55-64	presence.of.adults.age.65.64.specific.enhanced
Presence of adults – 45-54	presence.of.adults.age.45.54.specific.enhanced
Presence of adults – 35-44	presence.of.adults.age.35.44.specific.enhanced
Presence of adults – 25-34	presence.of.adults.age.25.34.specific.enhanced
Presence of adults – 18-24	presence.of.adults.age.18.24.specific.enhanced
Presence of children – 0-2	presence.of.children.age.00.02.specific.enhanced
Presence of children – 3-5	presence.of.children.age.03.05.specific.enhanced
Presence of children – 6-10	presence.of.children.age.06.10.specific.enhanced
Presence of children – 11-15	presence.of.children.age.11.15.specific.enhanced
Presence of children – 16-17	presence.of.children.age.16.17.specific.enhanced
Number of children	tsp.number.of.children.enhanced
Presence of one child	oneChild
Presence of multiple children	onePlusChild
Living area square footage	tsp.living.area.square.feet
Property lot size in acres	tsp.property.lot.size.in.acres
Age of home	ageOfHome
Dwelling type	tsp.advantage.dwelling.type
Income group (three categories)	lowIncome, medIncome, highIncome
Income group (five categories)	incomeGroupsA, incomeGroupsB, incomeGroupsC, incomeGroupsD, incomeGroupsE
Household income	advantage.household.income.legacy.dollars
Household education level	tsp.advantage.household.education
Home ownership status	tsp.advantage.home.owner
Number of adults	tsp.advantage.number.of.adults
Marital status	tsp.advantage.household.marital.status
Household age	tsp.advantage.household.age.enhanced
Length of residence	tsp.advantage.length.of.residence
Occupation	occupation
Household size	tsp.advantage.household.size.enhanced
Family composition	tsp.family.composition.enhanced
Trip count (all trips in 2013)	tripCount

Table 66 lists the neighborhood characteristic variables present in the marketing data set and used in the hierarchical tree-based regression analyses. The first six variables indicate the proportion of households in the given neighborhood with cars, trucks, motorcycles, and motor homes. The remaining variables indicate the average values of those vehicles.

Table 66: Neighborhood Characteristic Variables

Neighborhood Characteristics	
Percent of households owning a passenger car	percent.of.households.owning.a.registered.passenger.car
Percent of households owning a new passenger car	percent.of.households.owning.a.registered.new.passenger.car
Percent of households owning a truck	percent.of.households.owning.a.registered.truck
Percent of households owning a new truck	percent.of.households.owning.a.registered.new.truck
Percent of households owning a motorcycle	percent.of.households.owning.a.registered.motorcycle
Percent of households owning a motor home	percent.of.households.owning.a.registered.motor.home
Average value for new and used vehicles	average.cmv.in.thousands.for.all.new.and.used.registered.vehicles
Average value for new and used cars	average.cmv.in.thousands.for.new.and.used.registered.cars
Average value for new and used trucks	average.cmv.in.thousands.for.all.new.and.used.registered.trucks

Finally, Table 67 lists the interaction terms added to the data set. The first of these divides the toll amount by the segment count to create a measure of toll rate per corridor segment. This is meant to stand in contrast with the unmodified toll rate, which indicates the full amount paid. The second, toll amount divided by income, was created to investigate whether users consider the toll within the context of their overall income when making lane choice decisions. Note that the resulting value is very small due to the difference in magnitude between a single toll amount and a household's annual income. The income divided by household size term was included to better represent potential behavioral differences between households of the same income level but different family sizes.

Table 67: Interaction Terms

Interaction Terms	
Toll / Segment Count	tollSegments.HOT
Toll Amount / Income	tollIncome.HOT
Income / Household Size	incomeHhSize

Finding Problematic Variables with Regression Trees and Random Forests

Another benefit of the regression tree analysis used here is its aid in identifying variables that may be highly correlated with the dependent variable. In this case, that means identifying factors in the lane choice dataset that were endogenous to the dependent lane choice. The coded distance variable, for example, used the locations of the SRTA HOT and GP gantries to calculate the distance of each corridor trip. However, the HOT and GP gantries are at different locations and cover different portions of the corridor. The GP gantries span approximately 88% of the length of the HOT gantries. This method of calculating corridor trip distance results in distance as a discrete variable, rather than continuous, as there is a finite number of combinations of distance values that arise from the different start and end gantry combinations. Because the HOT and GP gantries are at different locations, the resulting distance value is highly correlated with the type of lane the vehicle trip is in. Only HOT trips extend the whole length of the corridor, for example. As a result, any corridor trip that exceeds roughly 13.5 miles must be a toll-lane trip. This will be illustrated below in Figure 137.

Initial regression tree experiments identified this and other variables that needed to be removed from future models. For example, a variable that calculated the speed difference between the Express Lanes and the general purpose lanes within the general purpose portion of a mixed trip only had values greater than zero during 'mixed' trips that used both lane types. Thus, any trip in which that variable had a value greater than zero was already known to have a toll lane component. The classification tree models identified this and the previously discussed

distance variables as the most powerful explanatory factors, but these are not causal variables and were then removed them from the models.

Figure 137 illustrates the regression tree results of the 2013 constructed trip data set, with paired Epsilon demographic data. In this run, both the morning and afternoon peak trips were pooled together. Note that the . The most striking element of the figure is the dominance of the 'distancemi' variable. The instances of this variable far outnumber any other factor in the tree; many branches contain multiple references to it. The problematic average speed difference variable discussed in the previous paragraph was identified in the random forest analysis; that discussion appears later in this chapter.

In the regression tree diagrams presented here, each node has three values underneath it. The two values in the first row represent the probability of each class at that node. In the case of the topmost node, the 0.47 value refers to the probability of a GP-only trip to the left, while the 0.53 value refers to the probability of a trip that includes an HOT portion to the right. This order is maintained across all nodes. The percentage value beneath the probabilities represents the observational shares at that node. Thus the topmost node includes 100% of all observations.

In this tree, the first split occurs on the distance variable. A trip distance value greater than 14 miles directs the user to the rightmost HOT node. Here the HOT option has a probability value of 100%; all of the trips include toll lane trips; none of them are GP-exclusive. This node represents 15% of the total trips in the data set. Trips under 14 miles in length make up 85% of the total observations. After the 14 mile distance slice, the next break point is again within the trip distance variable. Here the tree checks whether the trip distance is less than 4.5 miles, and then again whether it is less than 2.3 miles. The distance variable appears an additional five times, for a total of eight appearances in this initial regression tree. The only other variables

included are the Jimmy Carter Boulevard dummy, the general purpose lane transponder count, and the toll lane density (transponder count divided by distance).

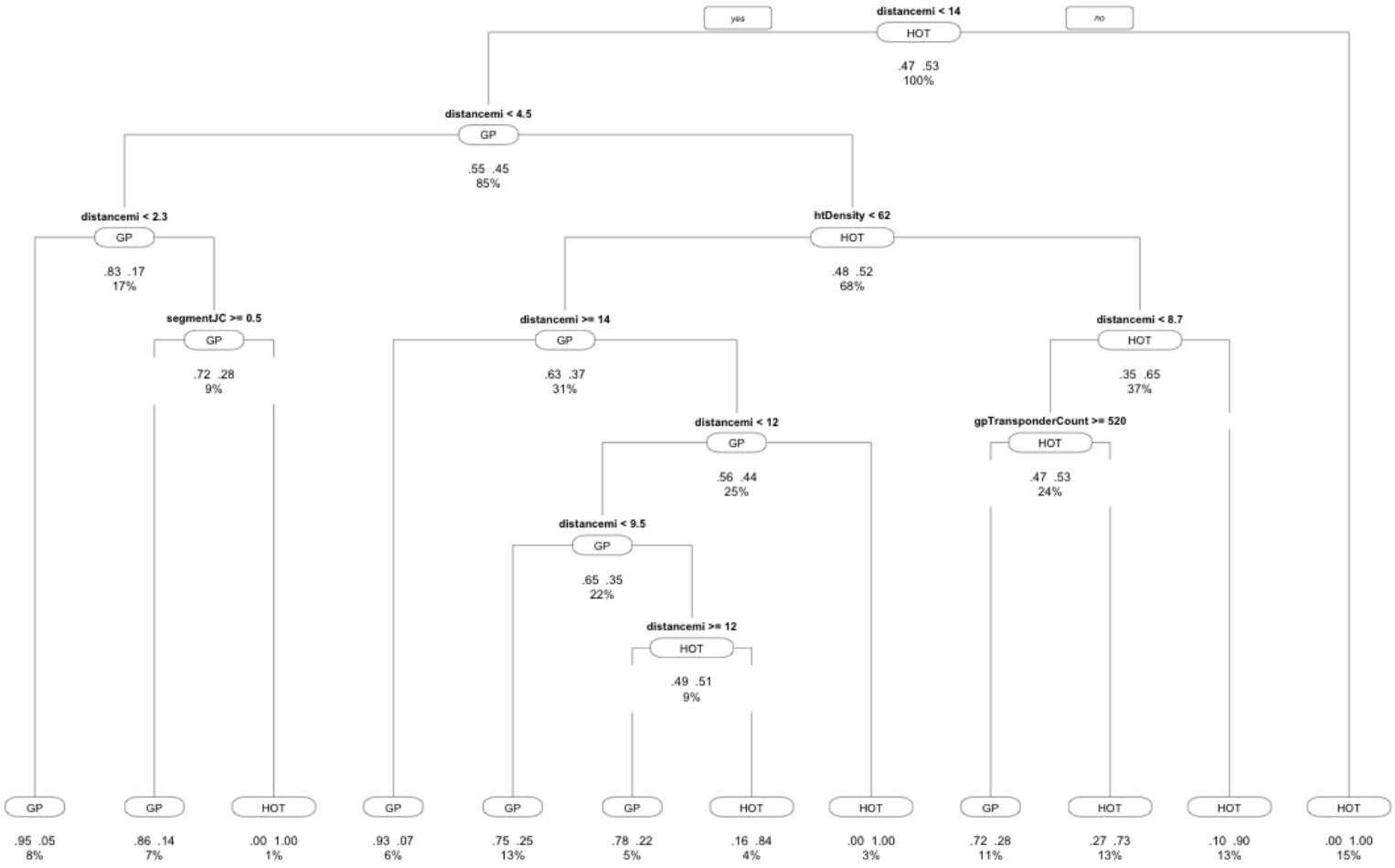


Figure 137: 2013 Regression Tree Results with Problematic Variables (n = 2,301,286)

Regression Tree Results without Problematic Variables

Performing the initial regression tree analysis helps visualize how significant an impact the 'distancemi' variable had on the toll lane models. The next step in the regression tree analysis was to remove that variable and re-run the models.

Figure 138 presents the regression results of the paired 2013 constructed trip data with both morning and afternoon trips pooled together and the 'distancemi' variable excluded from the tree design. The only common factor between this tree and the previous tree is the 'htDensity' variable, which counts the detected transponders per mile in the Express Lanes along the length of the user's trip. This new tree, which excludes the distance variable, no longer includes the raw toll amount variable and the two corridor segment dummy variables (indicating the presence of the vehicle in the Jimmy Carter Boulevard or Pleasant Hill Road segments of the HOT or GP lanes). This revised tree adds the segment count variable, which sums up the individual segment dummy results, the maximum daily toll amount (maxToll), and the average HOT and GP lane speeds. In the regression tree figures presented here, the percentage value listed underneath each 'leaf' show the shares of the observations that fall into that leaf.

The prominence of the segmentCount variable in this tree resembles the position of the distance variable in the previous problematic tree diagram. The segmentCount variable was coded to replace the distance variable; it avoids using the specific locations of the HOT and GP gantries. This replacement, while it doesn't appear in the tree as frequently as the distance variable did, is still positioned as the most impactful variable. Nearly a third of the total trips are partitioned out at the first level of the tree, which checks whether or not the segmentCount is equal to five (representing all five corridor segments).

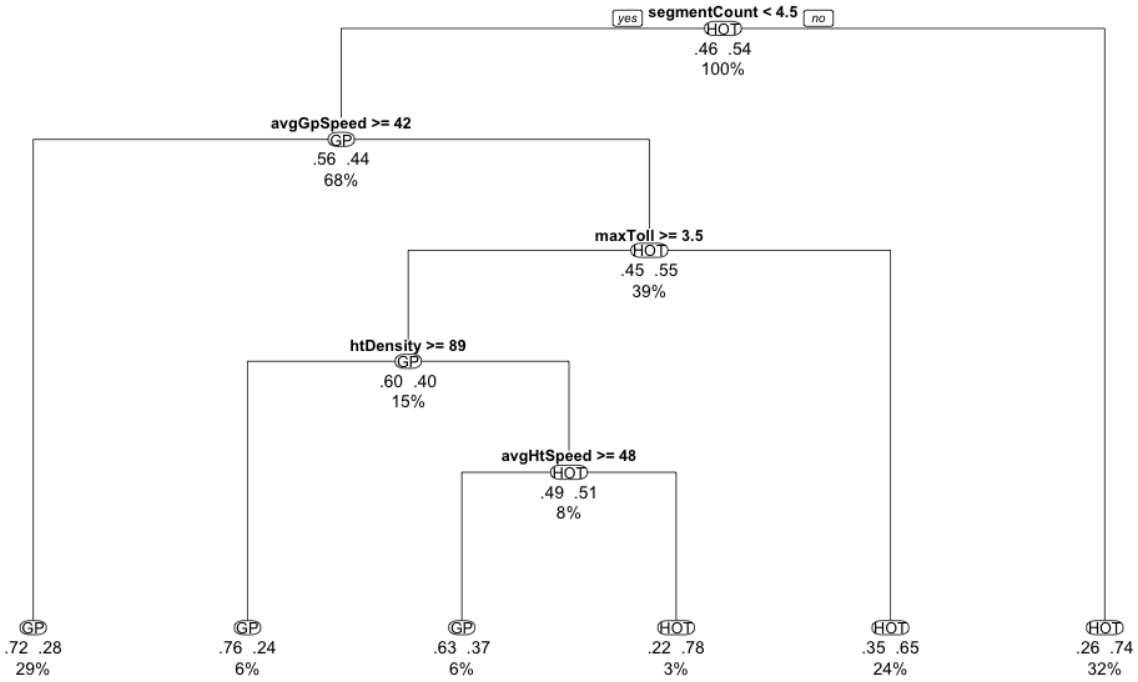


Figure 138: 2013 Pooled Regression Tree Results Without Problematic Variables (n = 2,301,286)

The next pair of figures presents the regression tree results after splitting the data set into the southbound morning peak trips and the northbound afternoon peak trips. Previous chapters, such as the Paired Versus Unpaired Data chapter, illustrated the benefits of such a split. This split also makes intuitive sense as morning and afternoon trips typically differ in their trip types and purposes. Figure 139 illustrates the regression tree results for the morning peak period southbound trips (1,107,026 trips in the data set). This model has substantial differences from the pooled tree; in particular, the segmentCount factor is less prominent for the AM southbound trips. More impactful is the vehicle's presence in the Old Peachtree Road segment; that dummy variable represents the first branch in the tree. Like the pooled tree, the AM tree identified the htDensity and maxToll variables as predictors of toll lane use. After the segmentOP variable, the most prominent factors are the maximum daily toll value and the average toll lane speed. The

segmentCount variable reappears at the third level of the tree, adjacent to the tollSegments interaction term and the toll lane transponder density.

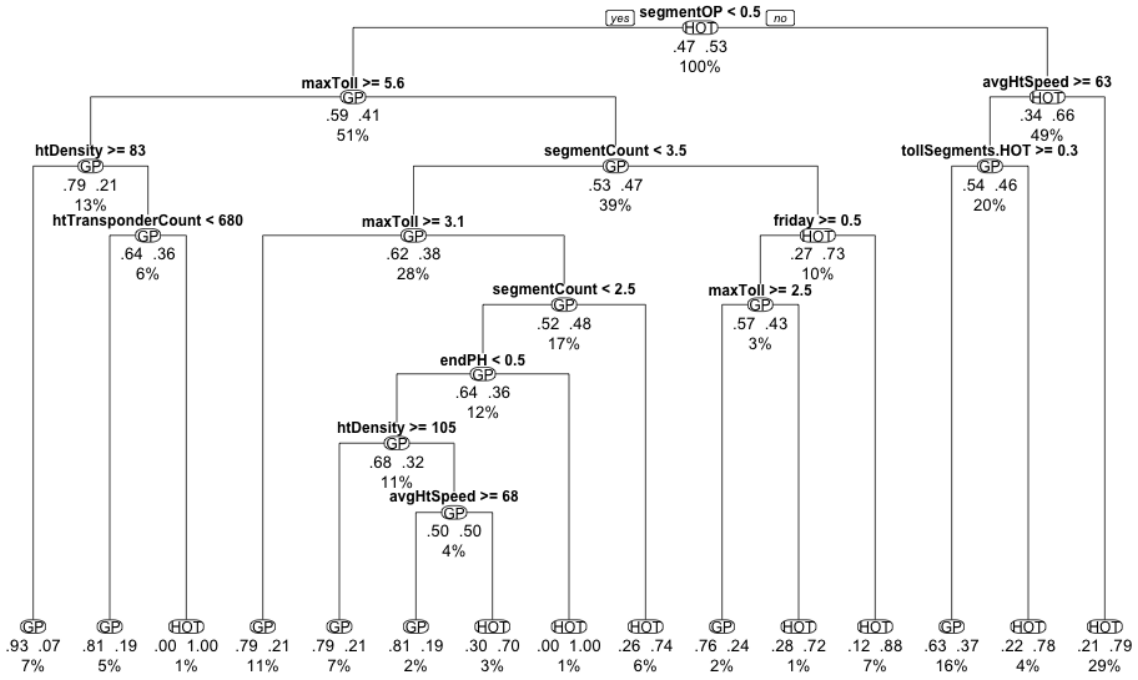


Figure 139: 2013 AM Peak Trips - Regression Tree Results (n = 1,107,026)

Figure 140 presented the regression tree results for the afternoon peak period northbound trips (1,194,260 observations). This tree more closely resembles that of the pooled data in its structure, particularly in its prioritization of the segmentCount variable, along with the second-level presence of the average general purpose lane speeds. The proportion of trips in the HOT node when the vehicle is detected on all five segments, 35%, is higher than that of the pooled tree. The segmentOP variable, indicating presence in the Old Peachtree Road segment of the corridor, appears here as well, though this factor applies to only 4% of the trips. Interestingly, that variable is in both the southbound AM and northbound PM tree results but not in the pooled tree model. The PM peak tree and the AM peak tree share four variables: along with the aforementioned segmentOP dummy variable, the maximum toll, segment count, and toll segment factors appear in both models. The PM tree omits the htDensity parameter which is included in the morning model, along with the average toll lane speed, toll lane transponder count variables, 'Friday' dummy indicator, and Pleasant Hill exit dummy indicator.

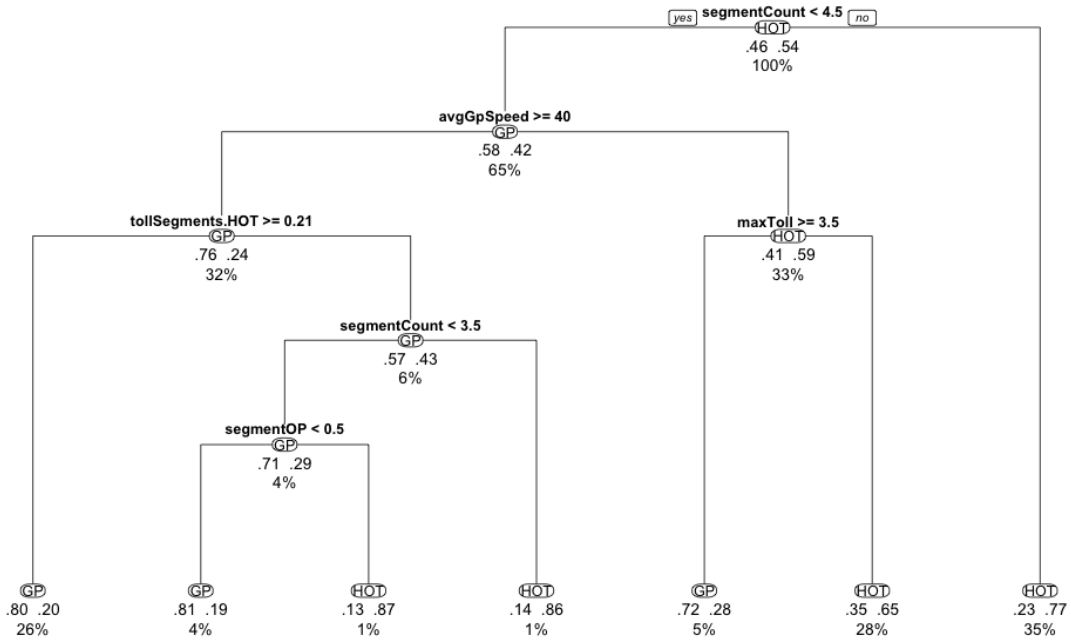


Figure 140: 2013 PM Peak Trips - Regression Tree Results (n = 1,194,260)

The results of the regression tree analysis for the 2013 constructed trips provide useful insights. In the three preceding models, covering the pooled data, AM-specific data, and PM-specific data, the only variables that appear in the regression trees are trip-specific and operational. The demographic factors provided by the marketing data set do not appear in these regression tree results. It may be the case that correlation with other variables results in the overshadowing of the demographic factors. The segmentCount variable (a stand in for trip distance) and the Old Peachtree Road dummy variable appear at the top of the three different tree diagrams. Toll rates appear in the form of the daily maximum value (maxToll) and interacted with the segment count (tollSegments, which divides the toll by the number of segments traversed), though not in their unmodified form (tollAmount.HOT). The average speed variables for both the toll lane and general purpose lanes are given greater prominence than the transponder count variables. Again, missing from these results is any sort of demographic factor. Despite the inclusion of 41 Epsilon household demographic variables, none of them had enough of an effect on HOT lane choice probability to appear in the regression tree results.

Regression Tree Results without Transponder Counts

An investigation into the correlation between the variables used in these hierarchical tree-based regression models, as well as in the initial modeling with in Chapter 8, provided valuable insights into the variable relationships. The results of this correlation analysis can be found in Appendix A. Among the primary findings from this analysis is the strong correlation between the transponder count variables and the speed variables. The toll lane transponder count variables were highly, and negatively correlated with the average GP lane speeds (correlation coefficient of -0.71) and the average HT lane speeds (correlation coefficient of -0.73). Toll lane

transponder counts were also positively correlated with the 50 mph congestion dummy (coefficient of 0.66) as well as the toll amount (0.84). The general purpose transponder counts were also negative correlated with average GP lane speeds (-0.16) and average toll lane speeds (-0.21), though the magnitudes were much lower. GP transponder counts were positively correlated with the congested50 indicator (0.21) and the toll rate (0.73). In the regression tree models presented above, average lane speed variables were more prominent than transponder counts. These finding led the author to remove the transponder count variables from the regression tree models and run them again to re-examine the results with fewer inter-correlations.

Figure 141 presents the morning peak period regression tree model minus the transponder count and htDensity variables. Again, perhaps the most striking result of this regression tree model is the lack of any demographic factors. This includes both the three category income factor and the five category income factor. The Old Peachtree Road indicator remains the first point at which the data are sliced. The removal of the transponder count variables (including with htDensity) does not result in new variables in the tree; in fact the indicator of trips ending at Pleasant Hill Road is also now absent from the model.

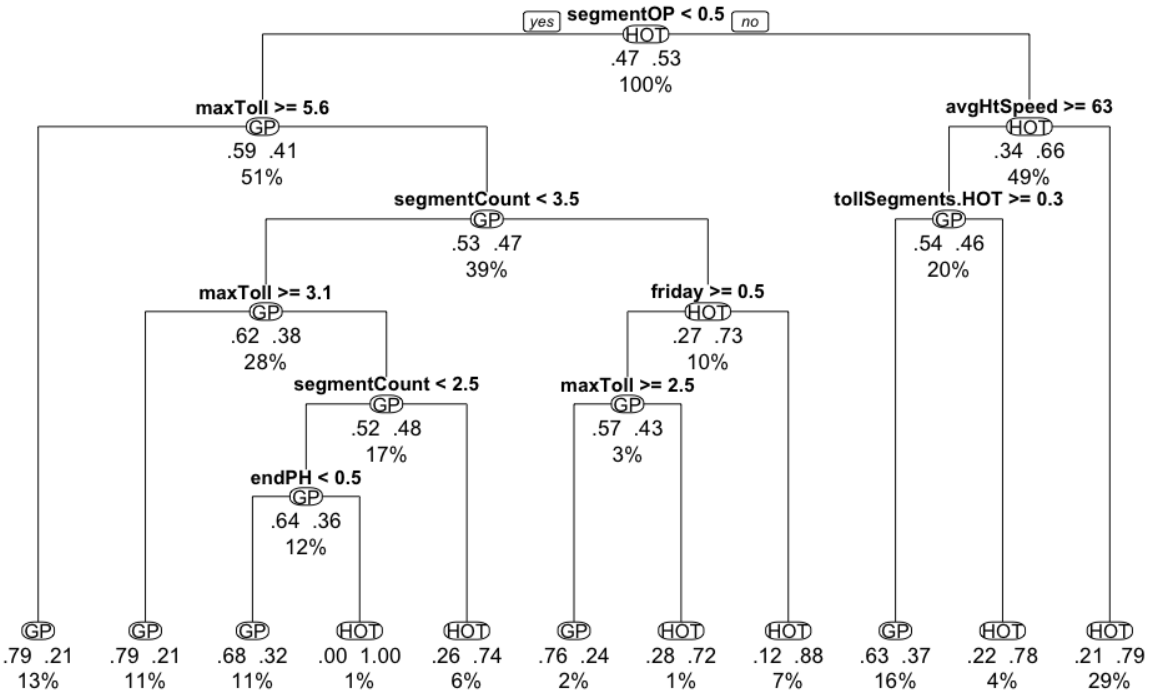


Figure 141: 2013 AM Peak Regression Tree Minus Transponder Counts (n = 1,107,026)

Figure 142 presents the afternoon peak period regression tree results minus the transponder count and htDensity variables. As was the case with the preceding models, all of the demographic factors are absent. The afternoon peak period tree shown in Figure 140 did not include any transponder count variables; the results displayed in this new tree are identical to those in the previous diagram.

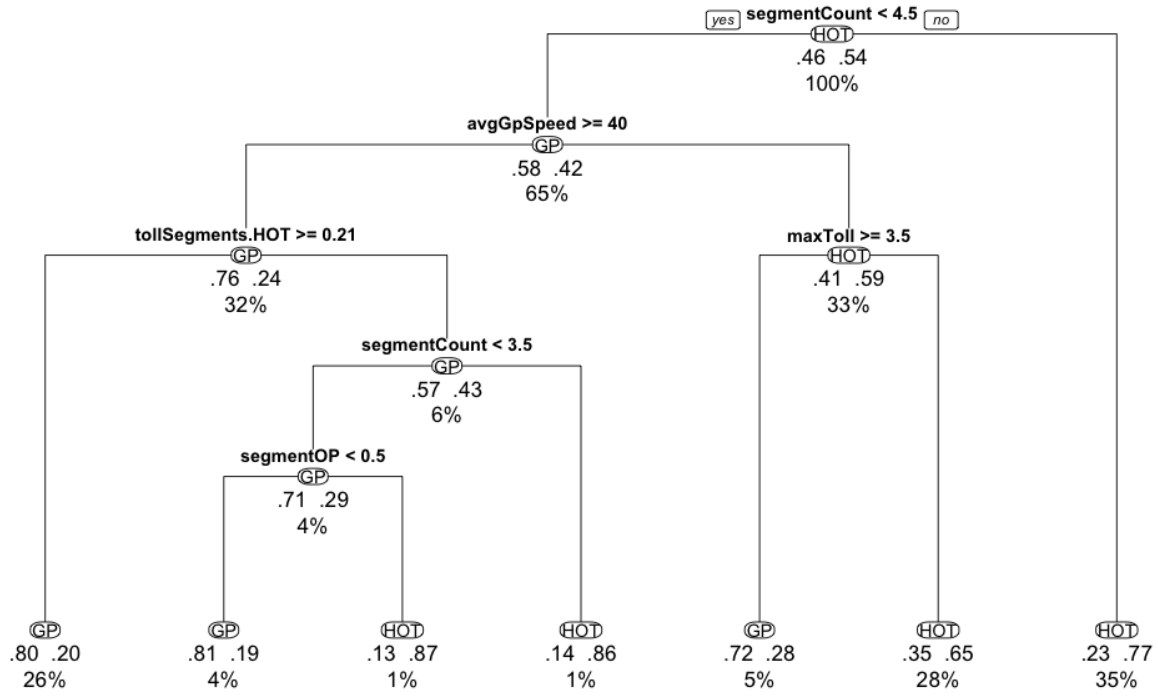


Figure 142: 2013 PM Peak Regression Tree Minus Transponder Counts (n = 1,194,260)

Random Forest Method

Much like the bootstrap analysis method, which uses repeated random sampling to construct confidence intervals, the random forest method extends the regression tree model through simulation. Here multiple regression tree models are estimated using subsamples of both the observations and the potential variables. This allows the power of variables that may be excluded from a single regression tree model to be estimated and interpreted (Breiman, 2001).

The random forest method is stricter than the regression tree method in that it is more sensitive to missing data. Random forests cannot be estimated with blank data values and as such will remove sample rows with any such values. The demographic data provided by Epsilon contains many variables with less than perfect coverage; that is, variables with at least some blank rows in the data set. The table below provides an overview of the variables in the 2013 paired trip data set that have blank rows. The proportion of blank rows is based on the total size of the weekday peak period and direction data set: 2,376,450 observations.

Table 68: Overview of Blank Rows in 2013 Trip Data

Variable Name	Number and Proportion of Blank Rows
presence.of.adults.age.18.24.specific.enhanced	**2314803 (97.4%)**
presence.of.adults.age.75.specific.enhanced	**2304623 (97.0%)**
presence.of.adults.age.65.74.specific.enhanced	**2229638 (93.8%)**
presence.of.children.age.00.02.enhanced	**2207316 (92.9%)**
presence.of.children.age.16.17.enhanced	**2205759 (92.8%)**
presence.of.children.age.03.05.enhanced	**2037399 (85.7%)**
presence.of.children.age.11.15.enhanced	**2028157 (85.3%)**
presence.of.children.age.06.10.enhanced	**2014093 (84.8%)**
presence.of.adults.age.25.34.specific.enhanced	**1989193 (83.7%)**
presence.of.adults.age.55.64.specific.enhanced	**1937215 (81.5%)**
presence.of.adults.unknown.age.enhanced	**1832071 (77.1%)**
presence.of.adults.age.35.44.specific.enhanced	**1723739 (72.5%)**
presence.of.adults.age.45.54.specific.enhanced	**1600730 (67.4%)**
tsp.number.of.children.enhanced	**1582445 (66.6%)**
tsp.property.lot.size.in.acres	**1069121 (45.0%)**
tsp.year.home.built.yyyy	**1001235 (42.1%)**
tsp.living.area.square.feet	**995676 (41.9%)**
tsp.family.composition.enhanced	122341 (5.1%)
occupation	103854 (4.4%)
tsp.advantage.home.owner	27587 (1.2%)
tsp.advantage.dwelling.type	14736 (0.6%)
advantage.household.income.legacy.dollars	4437 (0.2%)
incomeHhSize (interaction term)	4437 (0.2%)
tollIncome (interaction term)	4437 (0.2%)
tsp.advantage.household.marital.status	4437 (0.2%)
tsp.advantage.number.of.adults	4437 (0.2%)
tsp.advantage.presence.of.children.enhanced	4437 (0.2%)
percent.of.households.owning.a.registered.passenger.car	208 (0.01%)
percent.of.households.owning.a.registered.new.passenger.car	208 (0.01%)
percent.of.households.owning.a.registered.truck	208 (0.01%)
percent.of.households.owning.a.registered.new.truck	208 (0.01%)
percent.of.households.owning.a.registered.motorcycle	208 (0.01%)
average.cmv.in.thousands.for.all.new.and.used.registered.vehicles	208 (0.01%)
average.cmv.in.thousands.for.new.and.used.registered.cars	208 (0.01%)
average.cmv.in.thousands.for.all.new.and.used.registered.trucks	208 (0.01%)
percent.of.households.owning.a.registered.motor.home	208 (0.01%)

To deal with these missing observations, researchers removed those variables with a substantial number of blank rows. These variables, bolded in the table above, had more than 40% of the total observation count missing. The resulting data set was then narrowed to only complete cases, yielding 1,375,215 observations. The 'number of children' variable was replaced by two dummy variables: an indicator of the presence of one child, and an indicator of the presence of multiple children.

To investigate the impact of this narrowing of the data set, the author performed another regression tree analysis on the remaining observations. Figure 143 illustrates the results of this analysis, with the southbound AM-peak trips and the northbound PM-peak trips pooled together. The results are virtually identical to those of Figure 138, in which the full 2.3 million observation pooled sample, with no variables excluded, formed the basis of the regression tree analysis. The same factors have been selected in both tree figures, and the probability values at each node differ only by a maximum of 0.02.

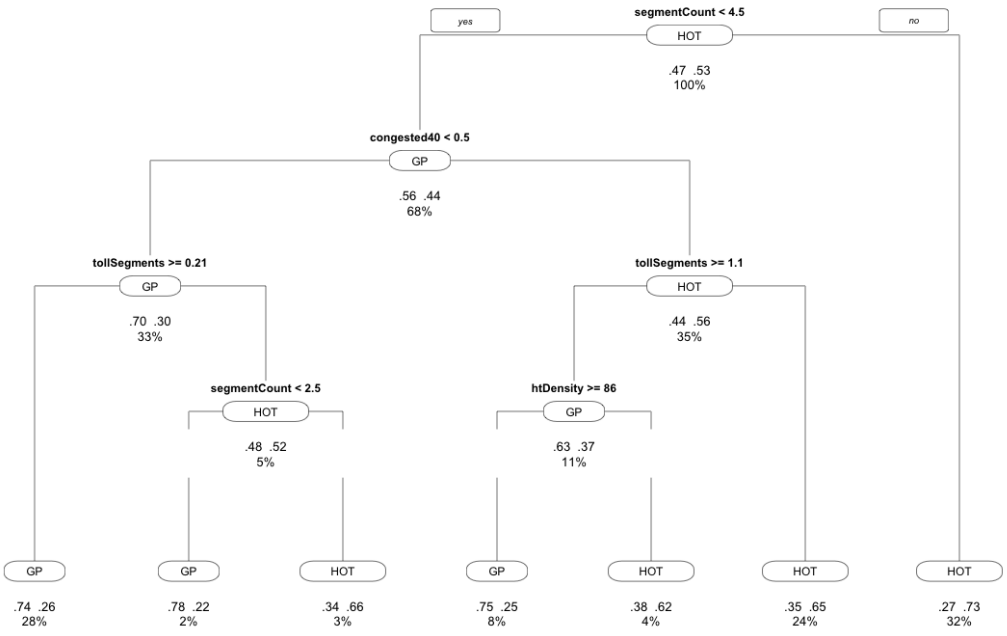


Figure 143: 2013 Regression Tree Results with Shortened Data Set (n = 1,331,604)

The random forest method was performed with 500 trees estimated and nine variables examined in each tree. The following figures present the results from the AM peak and PM peak random forest analyses. In each case the method generates two separate measures: variable importance and gini importance. Variable importance refers to the extent to which that variable affects model prediction error. The gini importance factor relates to the 'purity' of the nodes that result from splits on that variable. In both cases, higher importance values indicate more explanatory power (Liaw, 2002; Breiman, 2001).

Figure 144 presents the ranked variable importance results from the AM peak period southbound trips. The variables are presented in the order of their impact on model accuracy: in this figure, the htDensity has the largest impact and is ranked first. Here a large gap is present between the first variable (htDensity) and the next (daily maximum toll amount in the Express Lanes). Other notable results include the higher priority given to the transponder counts of the two lane types compared with the average speeds. The most significant congestion dummy variable is that which is activated at 50 miles per hour, followed closely by the congested45 dummy. The segment path that represents the full corridor, seg1to5 (Old Peachtree to I-285), is the highest path indicator to appear, followed by the seg2to5 indicator. By this point in the list, however, the impact on model accuracy is close to zero. As was the case in the regression tree results, no demographic variables appear in the top twenty random forest variable importance results.

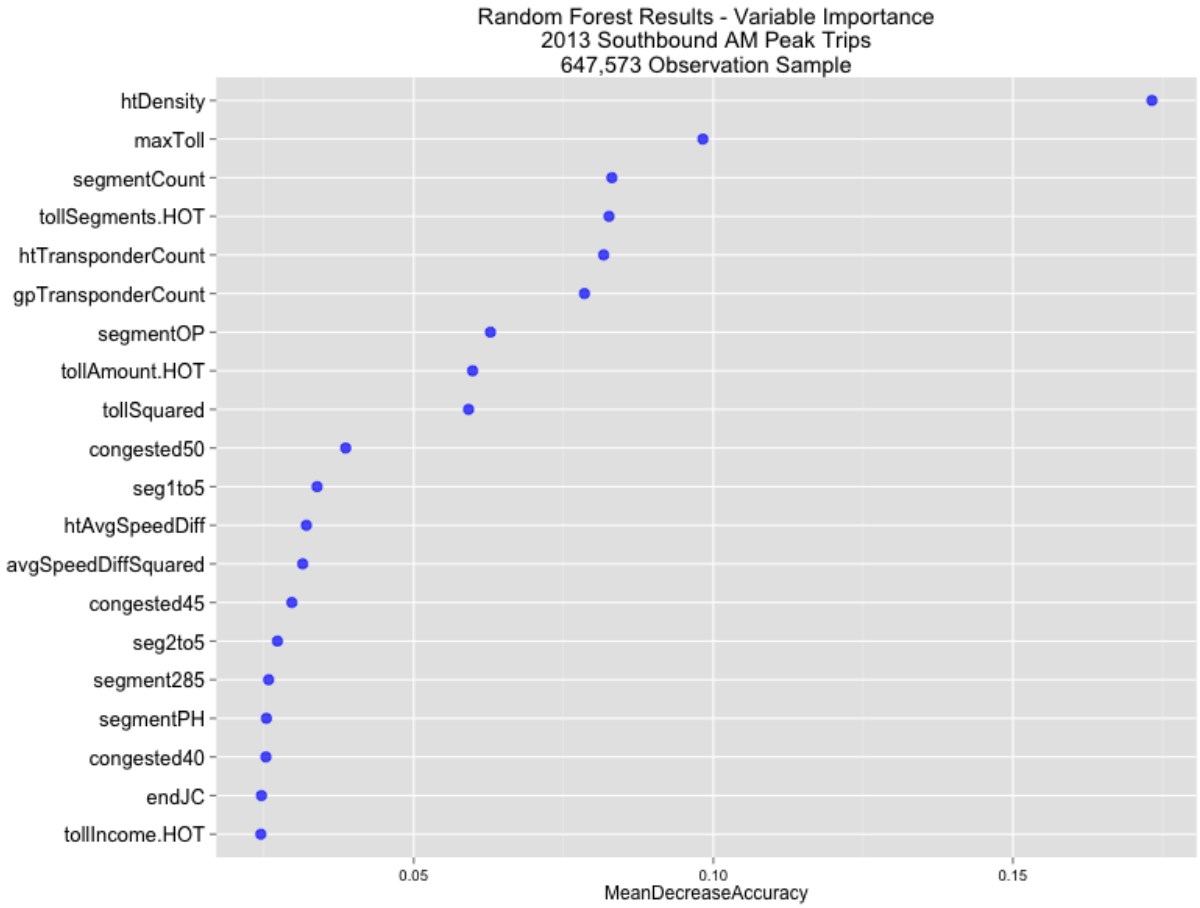


Figure 144: 2013 Southbound AM Random Forest Results - Variable Importance (n = 647,573)

Figure 145 illustrates the Gini Importance results from the southbound peak trips. As mentioned above, the Gini coefficient measures the impact each variable has on the homogeneity of its descendent nodes. Those variables that yield the most homogenous nodes have the largest decrease in their Gini coefficients (Dinsdale, 2013). This alternate measure provides another metric for model variable selection. In this case, the most impactful variable is once again the fifteen-minute measure of toll lane transponder counts per mile (htDensity). The next two variables are toll related: maximum daily toll, and toll divided by segment count. One substantial difference between this figure and the previous variable importance figure is the presence of marketing demographic factors. The most prominent of these are the neighborhood factors describing automobile, motorcycle, and truck ownership. Household income appears at the 20th position in the figure, interacted with household size in this case. Though these marketing demographic factors do appear here, compared with the previous regression tree models, their impact on node homogeneity is low.

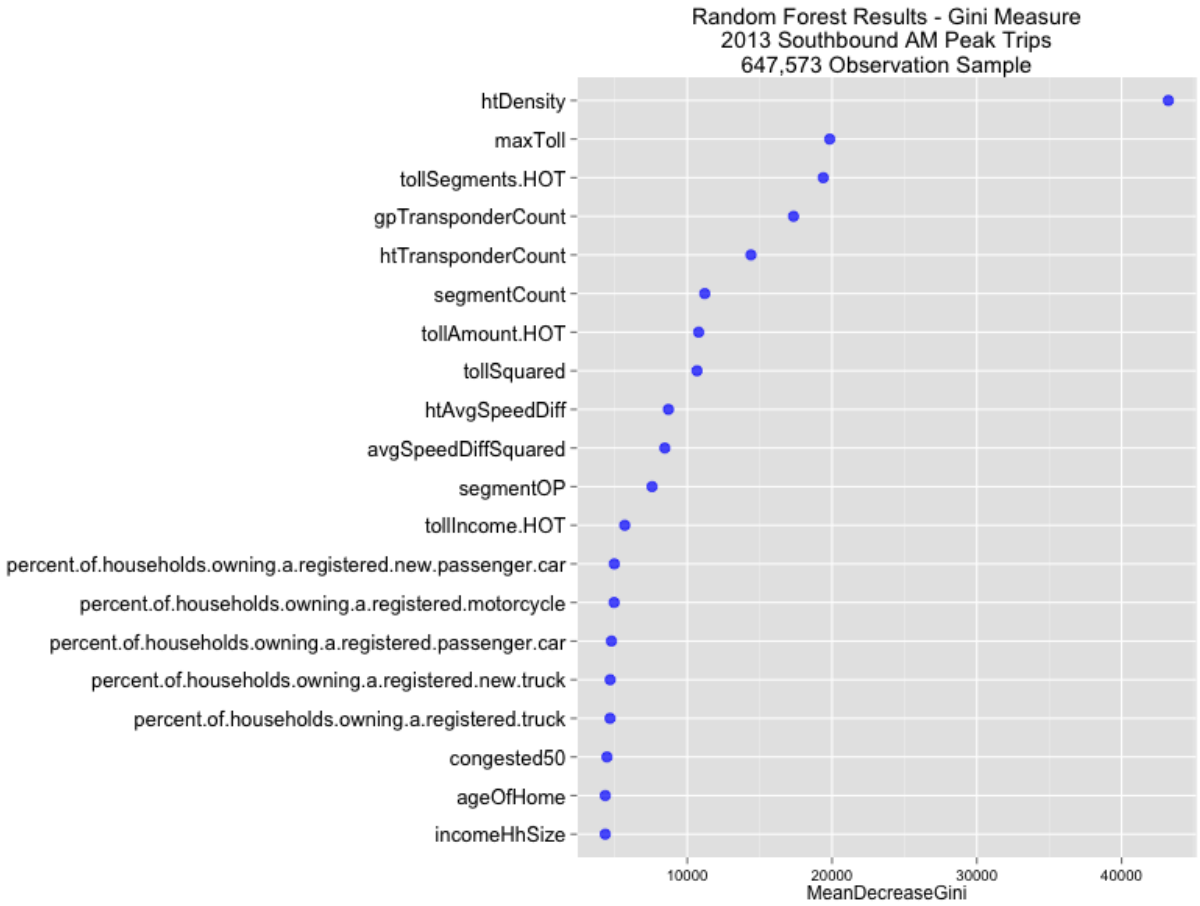


Figure 145: 2013 Southbound AM Random Forest Results – Gini Importance (n = 647,573)

Figure 146 and Figure 147 present the corresponding charts for the PM peak northbound trips. The most striking difference between the northbound variable importance chart and its southbound counterpart is the placement of the segmentCount factor at the top. In the afternoon trips, the previously-dominant htDensity variable is only marginally more impactful than the daily maximum toll term. As in the morning peak period results, the maximum toll is followed by the tollSegments interaction term (the toll amount divided by the total segment count, including both HOT and GP lane types). The corridor segment dummy variables are much more prominent in these afternoon trips: four of the five appear in the top twelve variables. The square of the toll amount is roughly equivalent with the toll amount itself. The afternoon trip models also benefit from the neighborhood-level auto ownership factors. The three congestion dummy variables that appear here, congested45, congested40, and congested35, are ranked closely together, but have little overall impact on model accuracy.

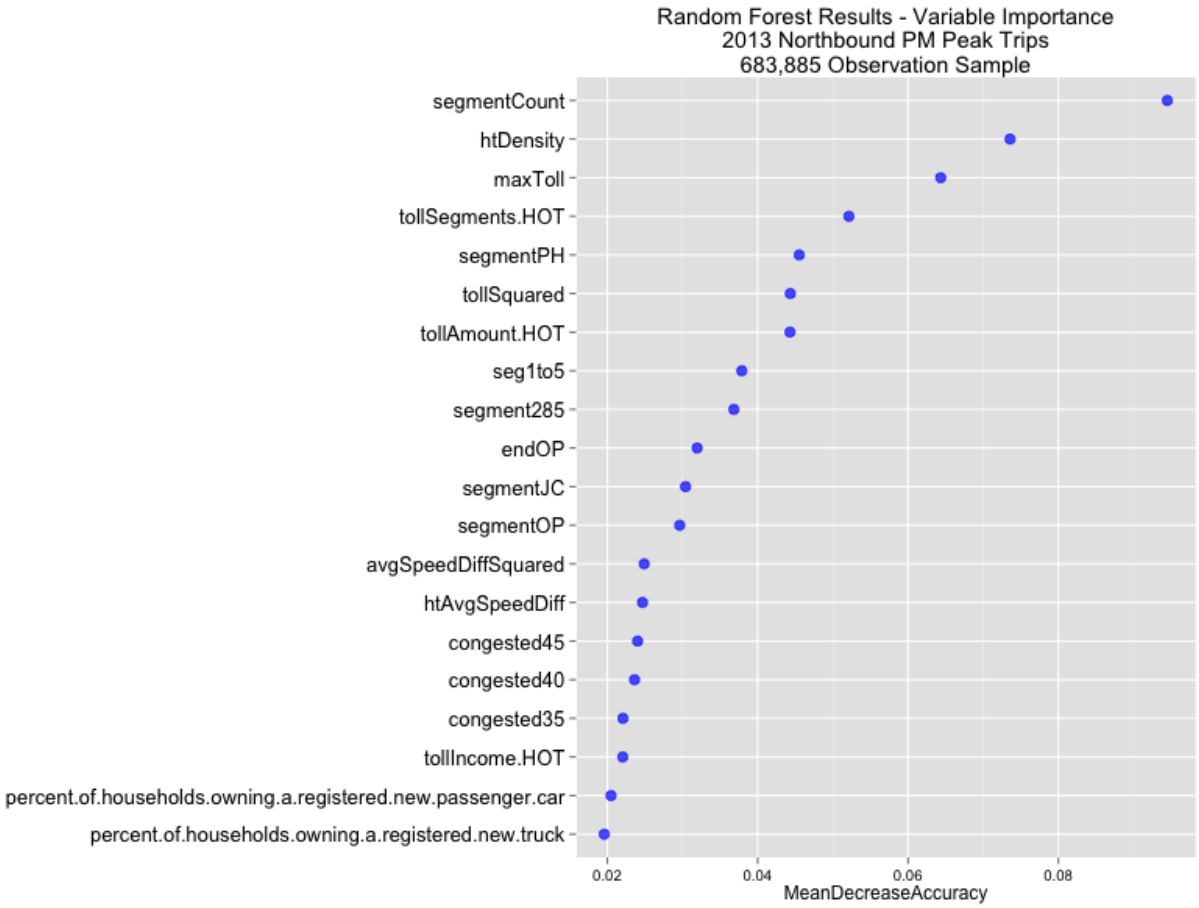


Figure 146: 2013 Northbound PM Random Forest Results - Variable Importance (n = 683,885)

Figure 147 shows the Gini importance results for the northbound 2013 trips. As in the variable importance chart, the top three factors are htDensity, segmentCount, and the daily maximum toll term. The next four factors include the toll amount and its square, the tollSegments interaction term, and the square of the lane type speed difference. The three GP congestion dummy variables that appear in the PM variable importance figure appear here as well; the neighborhood auto ownership characteristics are also similar. The household's age of home appears here and in the previous AM Gini importance chart, though not in either of the model accuracy charts. The only segment indicator that appears here is that of the Pleasant Hill corridor segment.

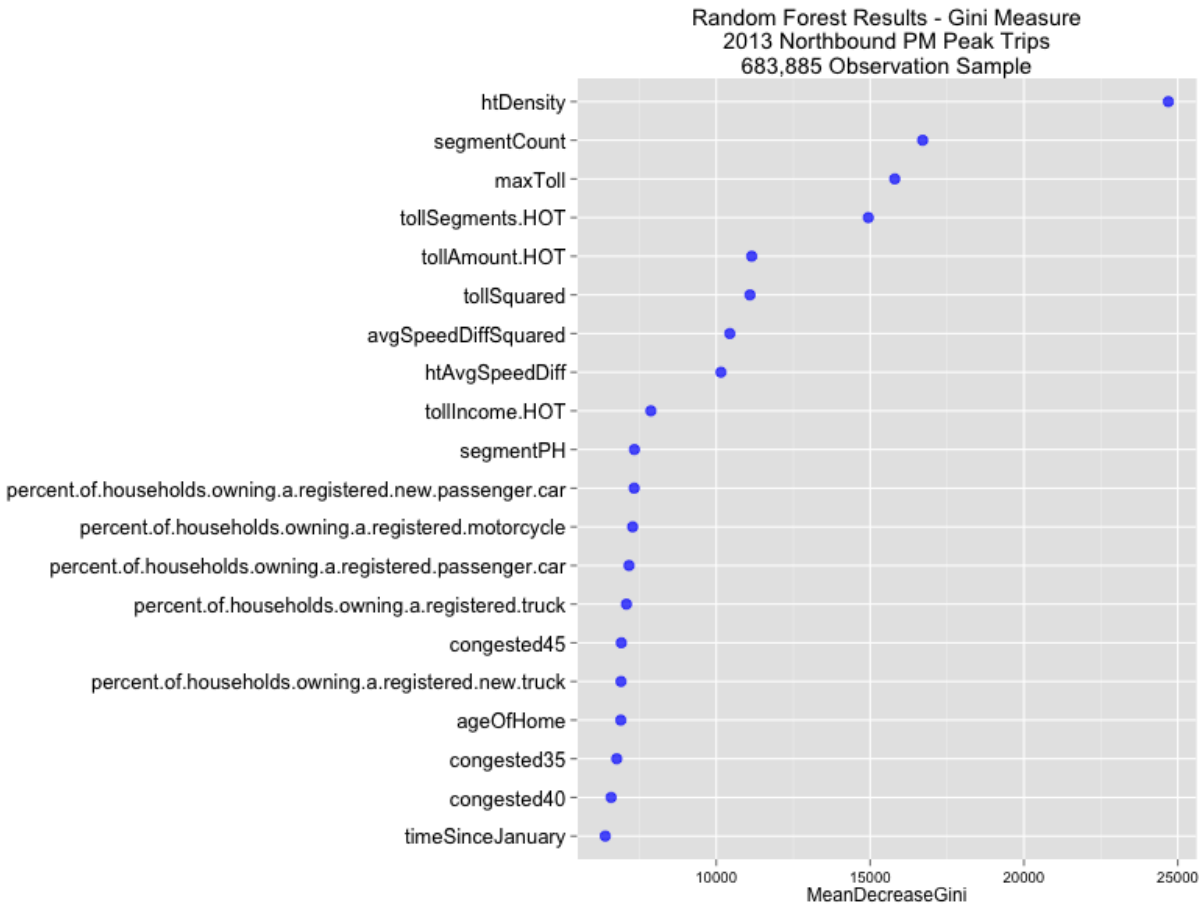


Figure 147: 2013 Northbound PM Random Forest Results – Gini Importance (n = 683,885)

Chapter Overview

The purpose of this chapter was to investigate the most impactful variables out of the hundred-plus factors available for the modeling analyses. The results indicated a strong preference for trip characteristics and operating conditions, with much less emphasis on household demographics. Across the various regression tree and random forest models, the most significant variables identified involved the length of the trip, variations on the toll amount, and the average speeds and transponder counts in both lane types.

The southbound and northbound trips yield different results in both the regression tree and random forest analyses. The most striking of these differences involves the lack of any transponder count variable in the PM peak regression tree results; these are represented in every other model through the htDensity, htTransponderCount, or gpTransponderCount variables. This perhaps speaks to the benefits of the bootstrapped random forest method over a single regression tree run. Other differences are more subtle, including different dummy variables indicating congestion at different GP speeds, and the relative ranks of the lane type speed difference and the square of that difference.

The most notable omissions from the regression tree results were the Epsilon demographic variables. In both the pooled model and the time/direction separated models, only the trip and operational characteristics appear in the tree diagrams. The toll variables that appear in the regression tree results, the daily maximum toll amount and the tollSegments interaction term, indicate a preference for Express Lane use at higher toll levels. That is, the regression trees select toll lane trips when the toll values are higher than the cutoffs. Though income does appear in the random forest results, it is never among the top five factors listed. Few household demographic factors besides income appear; household size appears in its interaction term with

income (income divided by household size), and age of home appears near the bottom of the morning and afternoon variable importance charts.

Also notable is the presence of neighborhood demographic variables in the random forest results. The inclusion of these variables was surprising as they are not household- or individual-specific, and toll lane decision making occurs at the individual level. The author suspects that these variables may be capturing some household attributes that are not present in the household or individual variables. While the marketing data provide household income information, for example, they do not provide overall wealth figures or household debts. This shortcoming is present in Census data as well. If households sort themselves into areas with similar financial characteristics, the neighborhood factors may help to provide more of a complete picture of household finances than income alone.

The regression tree and random forest methods come with their own limitations. Neither of them investigate correlation between the included factors. Note how both htDensity (based on the toll lane transponder count) and the htTransponderCount variables both appear in some random forest results. Similarly, lane type speed difference and toll amount both appear alongside the squares of those values. The insights they provide are still valuable, however, and the results of this chapter will be used to expand the initial lane choice models for the Modeling Extensions chapter.

CHAPTER 12

EXTENSION OF INITIAL MODELING ANALYSIS

Initial modeling efforts described in Chapter 8 had many avenues for improvement, even within the realm of basic binary logit modeling. Those avenues were described in detail in the Limitations section of that chapter. This chapter describes the improvements to the models that were incorporated after final publication of the paper. The chapter begins with an overview of the expanded data set used in the extended modeling analysis. The next section describes the various methodological changes and improvements, beginning with the additional variables and interaction terms considered in the models. The chapter then discusses the model building and variable selection strategies that the author used, as well as the mixed logit framework that replaces the standard logit framework. The next section presents the results from the model exploration. The chapter then discusses the results from the mixed logit models, and ends with an overview of the elasticity results.

Data

One major limitation of the initial modeling work was the scope of the data used in the analysis. While it did examine a full year's worth of constructed trips paired with household demographics, the full data set contains more than two additional years' of data. Similarly, the initial modeling work did not incorporate mixed trips, those that occur across both lane types. Further narrowing the sample was the low match rate between the SRTA Peach Pass transponders and the Epsilon demographic data due to the structure of the SRTA Account database. The many-to-many join issue between registered transponders and vehicles excluded accounts with more than one instance of either record type, so that only accounts with a single

vehicle and a single transponder remained. In addition to reducing the volume of data available for the analysis, this restriction also biased the results by including only households with a single registered vehicle.

Expanded Constructed Trip Data Set

This chapter expands the scope of the data under examination by rectifying many of the issues described above. In particular, the 'mixed' trips that were previously excluded from the analysis are now included, as are trips by SRTA accounts with multiple transponders. Table 69 provides an overview of the expanded data set for 2013, with the AM and PM peak trips separated. The afternoon period trips include more households, transponders, and observations. The demographic characteristics of both sets of users are similar, with median differences of only one unit in the household size and education categories. For the purposes of this dissertation, only trips from calendar year 2013 are included to save model estimation time.

Table 69: Expanded 2013 Data Overview

	Full Dataset – 2013	AM Peak Trips	PM Peak Trips
Unique Households Analyzed	36,854	27,774	33,482
Unique Transponders Analyzed	68,325	47,184	58,122
Total Trips Monitored	2,656,430	1,177,014	1,194,999
HOT-Exclusive Trips	386,370	205,765	168,309
GP-Exclusive Trips	1,337,286	558,477	547,978
Mixed Trips	932,774	412,772	478,712
Average Trip Speed (mph)	53.3	49.5	53.2
Average Segment Count	3.7	3.7	3.7
Median Household Income [25%; 75%]	$62,500 [$45,000; $112,500]	$62,500 [$45,000; $112,500]	$62,500 [$45,000; $112,500]
Median Household Size [25%; 75%]	3 [2; 5]	4 [2; 5]	3 [2; 5]
Median Household Age [25%; 75%]	5 [4; 6]	5 [4; 6]	5 [4; 6]
Median Household Education [25%; 75%]	5 [4; 5]	4 [4; 5]	5 [4; 5]

Table 70 presents the 2013 constructed trip data with income segment divisions. This table resembles that of the initial analysis in Chapter 8, with the inclusion of the previously excluded trips. While the Higher income segment has the lowest rate of GP-exclusive trips, and thus higher rates of HOT use, that segment's share of total Express Lane trips is lower than its share of households. This is likely due to differences in trips per household across the three segments.

Table 70: Expanded 2013 Data Overview - Income Segments

	Full Dataset	Lower Income ($0-50k)	Medium Income ($50-100k)	Higher Income ($100k+)
Unique Households Analyzed	36,854	10,127	15,588	11,139
% of Households by Income	100%	27.5%	42.3%	30.2%
Unique Transponders Analyzed	68,325	19,424	28,907	20,032
Total Trips Monitored	2,656,430	780,364	1,206,121	669,945
HOT-Exclusive Trips	386,370	113,915	167,577	104,878
GP-Exclusive Trips	1,337,286	409,743	610,314	317,229
Mixed Trips	932,774	256,706	428,230	247,838
% of HOT-Exclusive Trips	14.6%	14.6%	13.9%	15.6%
% of GP-Exclusive Trips	50.3%	52.5%	50.6%	47.4%
% of Mixed Trips	35.1%	32.9%	35.5%	37.0%
% of Total Trips by Income		29.4%	45.4%	25.2%
% of HOT Trips by Income		29.5%	43.4%	27.1%
% of GP Trips by Income		30.7%	45.6%	23.7%
% of Mixed Trips by Income		27.5%	45.9%	26.6%
Average Trip Speed (mph)	53.3	53.0	53.5	53.4
Average Segment Count	3.7	3.5	3.7	3.8

Methodology

The preliminary modeling work in the previous chapter had a number of methodological shortcomings that this chapter seeks to correct. The distance variable, calculated via the locations of the first and last detected gantries, was highly correlated with the lane type choice for reasons described earlier. The models used a limited set of independent factors, failing to incorporate any interaction terms or time-of-trip variables, among others. Repeated observations by the same transponders and households yielded serial correlation that resulted in biased estimators. Other factors, such as the congested dummy variable and the income segments, were used without exploring whether they were defined in the most appropriate manner.

Additional Variables and Interaction Terms

The first step in the exploratory process was to investigate the available variables that were not used in the initial models, and to examine interaction terms that may better describe user behavior relative to the individual terms. Table 71 presents the set of these additional variables

and interaction terms that were used to supplement the original data set. The 'congested' dummy variable used in the initial analysis was pre-defined without performing any sensitivity analysis. General purpose lanes may be 'congested' when average speeds are under 40 miles per hour; they may also be said to be congested under 30 mph. For this expanded modeling analysis, the author coded a series of dummy variables ranging from 5 miles per hour to 50 miles per hour in 5mph increments. The regression tree and random forest methods described in Chapter 11 identified the most impactful of these dummy variables, though those results varied across the different models. The other additional variables include dummy variables that indicate the time at which the trip was started at various levels: month, season, day of week, half-hour time interval, and hour-long time interval. The htDensity factor divides the transponder count in the toll lane by the length of the HOT segment traversed; note that this differs from the strict traffic engineering definition of density as the counts occur over a fifteen minute duration.

The interaction terms in the table include toll amount divided by household income, toll amount divided by number of segments, and income divided by household size. Two terms square the toll rate and lane type speed difference to examine whether non-linear effects better represent those factors. Note that the data set was restricted to include only trips where this speed difference is positive. The initial models pooled together the southbound trips in the AM peak period and the northbound trips in the PM peak period. The investigation of the paired and unpaired trip data examined northbound and southbound trips separately, which resulted in better model results in all cases. This chapter continues using that method.

Table 71: Additional Variables and Interaction Terms

Variable	Description
congested50 through congested05	Variables indicating average GP lane speeds
Month dummy variables	Month in which trip was taken
Season dummy variables	Season in which trip was taken
Day of week dummy variables	Day of week on which trip was taken
am600 – am930	Dummy indicating half-hour interval for trip start time
pm1500 – pm1830	Dummy indicating half-hour interval for trip start time
sixAm – nineAm	Dummy indicating hour long interval for trip start time
threePm – sixPm	Dummy indicating hour long interval for trip start time
htDensity	Transponders per mile in HOT lane, 15 minute count
segmentCount	Number of segments traversed
tollIncome	Toll divided by log of household income
tollSegments	Toll divided by number of segments
incomeHhSize	Income divided by household size
Income Segments: 5 groups	Dummy variables indicating presence in one of five income segments
tollRateSquared	Square of toll amount
avgSpeedDiffSquared	Square of average speed difference between HOT and GP lanes

Alternative Income Segmentation Investigation

One of the main takeaways from the initial modeling work and the value of travel time savings work was the behavioral similarity among the three pre-defined income segments. The boundaries of these segments were selected based on the number of households that fell into each income category; the purpose of the selected intervals was to make the household counts similar in each segment. This chapter extends the initial analysis by further segmenting the 'Higher' income group (households with over $100k in annual income) into smaller partitions to investigate potential variability in lane choice determinants for those households. The motivation behind this is to examine whether users highest end of the income spectrum exhibit different decision making processes than those closer to the Medium/Higher income boundary.

Model Building Strategy

The model building process employed in this chapter takes an iterative approach to adding new variables. After examining the impact of the new factors and selecting a new base model, the

interaction terms are added to the resulting model by themselves and in various combinations. The author used the Akaike Information Criterion (AIC) measure to compare models of different parameter counts to investigate whether the benefit of additional variables outweighed the cost of their inclusion. The following table lists and describes the models and the progression of the model building strategy.

Table 72: Model Numbers and Descriptions

Model Number	Description
Model 1	Recreation of Initial TRB Model (Sheikh, 2015)
Model 2	Distance variable replaced with segment counts (HOT or GP)
Model 3	Average speed difference replaced with square of average speed difference (positive differences only)
Model 4a, 4b, 4c	Comparison of congestion dummy variables
Model 5a, 5b	Incorporation of selected congestion dummy variables
Model 6a, 6b	Added month and season dummy variables
Model 7	Added day of week dummy variables
Model 8	Added trip start time hour dummy variables
Model 9	Replaced trip start time hour dummy variables with half-hour indicators
Model 10	Replaced toll variable with square of toll
Model 11a, 11b	Replaced transponder counts with htDensity
Model 12a, 12b	Added tollLogIncome interaction term (toll divided by log(income))
Model 13	Added tollIncome interaction term (toll divided by income)
Model 14	Added income and log(income) divided by household size interaction terms
Model 15	Added toll divided by segment count interaction term
Model 16	All interaction terms included
Model 17	Additional interaction term combinations

Mixed Logit Modeling

As discussed in the Initial HOT Use Choice Analysis chapter, the standard binary logit framework has certain limitations that can affect the modeling results. Most relevant among these for this analysis is the issue of serial correlation, or repeated choices by the same individuals. The standard logit framework assumes independence among the errors in the model; this assumption is violated when the same user appears multiple times in the data set. Repeated observations are a defining characteristic of the combined SRTA-Epsilon lane use data set, and as such any standard logit model estimated on that data will be biased due to serial correlation.

The estimators provided by the standard binary logit framework are also fixed, and may not represent the range of responses to a specific variable. The independence from irrelevant alternatives (IIA) property is typically a limitation as well, but in this binary framework there are only two alternatives (Train, 1986).

To address these issues, the author used the models selected from the standard binary logit analysis to estimate mixed logit models of the same design. The mixed logit framework addresses the serial correlation issue by identifying the user making each choice and adjusting error terms appropriately. Mixed logit models also allow for random parameters: the modeler specifies a distribution for a parameter, and the model estimates a random variable coefficient to fit that distribution. This allows the models to better represent the potential variation in user response to different factors. For this analysis, all of the models were estimated using 500 Halton draws for the simulation. This chapter will present the results of these mixed logit models, along with the parameter distributions that arise from each one. Note that for some of the model runs, a random sub-sample of 10,000 records was used. This restriction was meant to save estimation time. For the model that was selected as the preferred mixed logit model, the full data sets were used in both the AM and PM peak periods.

Modeling Results

Previous Model with Expanded Data Set

The extended modeling analysis began by re-estimating the initial models with the newly expanded data set. Table 73 presents the results from the model used in the Initial Use Choice Modeling chapter and the Sheikh, 2015 TRB paper. This table incorporates the additional mixed-lane-type trips from 2013; the dependent variable is use of the priced facility for any portion of the trip. Trips which use both lane types are modeled as HOT trips; the corridor

conditions used in their records compare the segments which the user traversed in the Express Lanes. In the model results tables below, the reported R^2 measure is McFadden's pseudo-R^2 value for discrete choice models. This measure measures the log likelihood of the full model against the log likelihood of the intercept-only model. For each model presented in this chapter, odds ratios were calculated from the estimated coefficients. An odds ratio represents the increase of the odds of an event given a unit increase in the independent variable for which it was estimated. In this research, the dependent event is the use of the Express Lanes for a portion of a trip. The odds represent the proportion of positive outcomes (HOT use) versus negative outcomes (no HOT use); the odds ratio estimates the increase in those odds (Szulimas, 2010). The full set of odds ratio results can be found in Appendix C.

The re-estimation of the initial model was performed on a much larger data set; over a million additional records were included. The timeframe for the trips, calendar year 2013, did not change. The count of unique Epsilon households increased by nearly 8,000 (27.3%). The results illustrate differences in coefficient signs and magnitudes among the pooled models. In particular, the average speed difference and transponder count estimators flip their signs: previously the avgSpeed coefficient was positive while the transponderCount coefficient was negative. The correlation matrices in Appendix A illustrate the high level of correlation between vehicle speeds and transponder counts, indicating the reason for this effect. In addition, the Paired versus Unpaired Data chapter identified a similar effect between the average speed difference and congested conditions factors. Once again, the distance variable yields the highest t-statistic; for reasons discussed previously, it was removed from all future models. The goodness of fit, as measured by the pseudo-R^2 value, improves with the additional observations. Like the previous pooled model, all of the coefficients are significant at the 95% confidence

level, and here most are significant at the 99% confidence level. Only the household education factor sees a decrease in its relative significance. As in the previous model, age, household size, and education all have negative coefficients. The Data Sources chapter investigated the correlation among these demographic factors and found that all three are positively correlated with household income.

The previous publication did not segment the trip set by peak period and direction; Table 73 shows the results of this segmentation. The differences here are in the goodness of fit, the average speed difference coefficient, and the household education coefficient. Segmenting the models yields a higher R^2 value for the afternoon peak, though the morning peak model sees a decrease relative to the original pooled model. The average speed difference coefficient is positive in the afternoon, resembling the results from the previous paper rather than the pooled and morning peak models in Table 73. Household education in only the morning peak has a positive effect on the probability of toll lane use; while it is significant at the 95% confidence level, the estimator is still close to zero in magnitude. Note that in all of the modeling results tables below, the 'HOT:' prefix indicates that the coefficient is alternative specific and the estimator reflects the change in probability of using the Express Lanes.

Table 73: Re-Estimation of Initial Model for TRB 2015

	Pooled	AM Peak – Model 1	PM Peak – Model 1
Intercept	-2.904*** (t = -96.062)	-2.236*** (t = -53.675)	-3.775*** (t = -85.844)
avgSpeed	-0.012*** (t = -48.708)	-0.030*** (t = -93.353)	0.016*** (t = 39.447)
tollAmount	-0.567*** (t = -443.468)	-0.558*** (t = -327.322)	-0.557*** (t = -264.996)
transponderCount	0.004*** (t = 245.240)	0.003*** (t = 158.524)	0.004*** (t = 179.556)
HOT: southbound	0.143*** (t = 36.443)		
HOT: congested40	1.488*** (t = 312.448)	1.469*** (t = 212.343)	1.367*** (t = 200.891)
HOT: log(hhIncomedollars)	0.043*** (t = 14.150)	0.041*** (t = 9.774)	0.062*** (t = 14.222)
HOT: hhEdu	-0.014*** (t = -5.945)	0.011*** (t = 3.418)	-0.064*** (t = -19.164)
HOT: hhAge	-0.020*** (t = -13.671)	-0.030*** (t = -14.649)	-0.013*** (t = -6.110)
HOT: hhSize	-0.039*** (t = -40.095)	-0.049*** (t = -36.282)	-0.029*** (t = -20.111)
HOT: distancemi	0.398*** (t = 671.968)	0.367*** (t = 444.069)	0.431*** (t = 490.117)
HOT Share	0.5355	0.5287	0.5418
Observations	2,297,048	1,105,171	1,191,877
R^2	0.261	0.223	0.307
Log Likelihood	-1,171,791.00	-594,075.40	-569,297.80

*p<0.1; **p<0.05; ***p<0.01

In the models in Table 73, the odds ratios of the four demographic factors are all close to one (the value that indicates that an increase in the factor value does not change the odds of the dependent event). The highest of those four is income with an odds ratio of 1.06 in the PM peak. The lowest, household education, has an odds ratio of 0.938 in the PM peak. Similarly, the odds ratio of the difference in transponder counts is 1.00. Of the remaining factors, the congested conditions dummy variables have the largest odds ratios (4.35 in the morning, 3.93 in the afternoon), and the toll amount odds ratios have substantial negative impacts (0.57 in both periods).

Additional Variables

The following series of models examine the impacts of adding additional variables listed in Table 71 to the initial model. After presenting all of the individual models, the goodness of fit and Akaike Information Criterion (AIC) values are summarized in Table 86. The author examined the AIC measure to investigate the benefit of each additional variable or set of dummy

variables relative to the cost of the additional factors. Those measures were used to select a model to use in further investigations.

Model 2 – Replacing Distance with Segment Count

Table 74 presents the results of replacing the problematic 'distancemi' variable, which measured the distance in miles between corridor gantries, with the 'segmentCount' variable, which counts the number of corridor segments (out of five) on which the vehicle was detected. This results in a large decrease in the pseudo-R^2 measure; the decrease exceeds 0.10 in both cases. Note that this decrease represents a correction in the model; the distance variable was not independent and as such had a large impact on the goodness-of-fit. The segmentCount coefficient remains positive and, like the removed distance coefficient, has the highest t-statistic of all of the factors. Note that in this and all successive tables, the new or changed variables are shaded in grey.

Table 74: Distance Replaced with segmentCount

	AM Peak – Model 2	PM Peak – Model 2
Intercept	-2.573*** (t = -67.052)	-3.911*** (t = -96.772)
avgSpeed	-0.010*** (t = -35.789)	0.013*** (t = 33.510)
tollAmount	-0.352*** (t = -241.986)	-0.340*** (t = -183.642)
transponderCount	0.0005*** (t = 27.748)	0.003*** (t = 135.881)
HOT: congested40	1.395*** (t = 220.169)	1.192*** (t = 192.650)
HOT: log(hhIncomedollars)	0.124*** (t = 32.380)	0.128*** (t = 32.089)
HOT: hhEdu	-0.096*** (t = -33.077)	-0.200*** (t = -65.274)
HOT: hhAge	-0.028*** (t = -15.457)	-0.006*** (t = -3.385)
HOT: hhSize	-0.044*** (t = -35.465)	-0.020*** (t = -15.236)
HOT: segmentCount	0.655*** (t = 286.978)	0.955*** (t = 396.364)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R^2	0.102	0.205
Log Likelihood	-686,021.40	-653,349.60

*p<0.1; **p<0.05; ***p<0.01

In the previous set of models (Model 1), the odds ratios of the distance variables were 1.44 in the morning peak and 1.54 in the afternoon peak. The segment count variable odds ratios are 1.93 in the morning and 2.60 in the afternoon. Note that while the range of distance values

extend from roughly one mile to fifteen miles, the segment count factor has a minimum value of one and a maximum value of five.

Model 3 – Square of Average Speed Difference

Table 75 shows the results of squaring the average speed difference factor. The motivation for this change was to investigate potential non-linear impacts of speed differences. The results suggest that this variant is preferable; model fit improves marginally, while the t-statistic increases in both the AM and PM peak models. The afternoon peak coefficient is now negative, like the morning peak coefficient and unlike the afternoon peak coefficient in the previous two models. This result is counterintuitive as it suggests that users are more likely to use the Express Lanes when they are slower than the GP lanes, though it agrees with the interpretation of the transponderCount coefficient: users are more likely to use the Express Lanes when they have more Peach Pass holding vehicles than the GP lanes. This is another case where the magnitudes are very close to zero, however.

Table 75: Square of Average Speed Difference

	AM Peak – Model 3	PM Peak – Model 3
Intercept	-2.429*** (t = -63.122)	-3.464*** (t = -85.773)
$avgSpeed^2$	-0.0002*** (t = -83.558)	-0.0002*** (t = -49.912)
tollAmount	-0.370*** (t = -250.724)	-0.362*** (t = -191.343)
transponderCount	0.0003*** (t = 18.684)	0.003*** (t = 133.269)
HOT: congested40	1.428*** (t = 237.063)	1.374*** (t = 252.782)
HOT: log(hhIncomedollars)	0.127*** (t = 33.208)	0.124*** (t = 31.049)
HOT: hhEdu	-0.103*** (t = -35.106)	-0.195*** (t = -63.643)
HOT: hhAge	-0.029*** (t = -15.778)	-0.006*** (t = -3.256)
HOT: hhSize	-0.045*** (t = -35.828)	-0.021*** (t = -16.263)
HOT: segmentCount	0.672*** (t = 291.415)	0.998*** (t = 405.242)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R2	0.106	0.206
Log Likelihood	-683,125.30	-652,655.80

*p<0.1; **p<0.05; ***p<0.01

While the goodness of fit and test statistics both improve, the odds ratio measures change very little. Model 2 saw odds ratios of 0.99 and 1.01 for the difference in average speed

measures for the AM and PM peaks respectively. The odds ratios of the squared speed differences in Model 3 are 1.00 in both cases.

Model 4 – Congestion Dummy Comparison

Table 76 and Table 77 show the results of six univariate models, examining three congestion dummy variables each in the morning and afternoon peak periods. The dummy variables (50, 45, and 35 mph in the morning; 45, 40, and 35 mph in the afternoon) were previously identified in various iterations of the regression tree models presented in Chapter 11. Of the ten different dummy variables, they were found to be the most impactful by the regression tree and random forest analyses. Table 76 presents the results of the AM peak univariate models. The three different congestion dummy variables are similar in sign and magnitude. The goodness of fit measures also differ only marginally, with the congested35 dummy yielding the lowest fit. Of the remaining two, the slightly increased significance of the congested50 dummy makes it the favored variable for the morning period models.

Table 76: AM Congestion Dummy Variable Comparison

	AM Peak Model 4a	AM Peak Model 4b	AM Peak Model 4c
Intercept	-0.510*** (t = -139.481)	-0.403*** (t = -124.720)	-0.185*** (t = -72.586)
HOT: congested50	0.878*** (t = 203.322)		
HOT: congested45		0.813*** (t = 200.938)	
HOT: congested35			0.698*** (t = 178.124)
HOT Share	0.5287	0.5287	0.5287
Observations	1,105,171	1,105,171	1,105,171
R^2	0.028	0.027	0.021
Log Likelihood	-742,933.20	-743,613.70	-748,076.60

*p<0.1; **p<0.05; ***p<0.01

Table 77 presents a similar comparison for the PM peak period. Here the model fits differ by larger amounts, with the 45 mph dummy yielding the best goodness of fit. The congested45 dummy also had the highest t-statistic, and the model had the highest log likelihood value. The regression tree results indicated that the congested40 variable was more impactful,

while the random forest method placed more importance on the congested45 variable. For this reason the next model iteration compares both dummies in the presence of the other factors.

Table 77: PM Congestion Dummy Variable Comparison

	PM Peak Model 4a	PM Peak Model 4b	PM Peak Model 4c
Intercept	-0.691*** (t = -215.318)	-0.462*** (t = -171.212)	-0.262*** (t = -111.791)
HOT: congested45	1.363*** (t = 338.402)		
HOT: congested40		1.256*** (t = 325.504)	
HOT: congested35			1.195*** (t = 294.594)
HOT Share	0.5418	0.5418	0.5418
Observations	1,191,877	1,191,877	1,191,877
R^2	0.074	0.068	0.056
Log Likelihood	-761,241.20	-766,362.40	-775,714.20

*p<0.1; **p<0.05; ***p<0.01

Model 5 – Incorporating Congestion Dummies in Models

Table 78 presents the results with the new congestion dummy variables that were identified through the regression tree and univariate modeling investigations. For the morning peak period model, the substitution of the congested50 variable for the previously used congested40 variable yields an increase in the model fit and the t-statistic for that factor. The signs and magnitudes of the other variables remain similar. The two afternoon peak models compare the two congestion dummy levels presented in Table 77. While the congested45 dummy performed better than congested40 in isolation, the opposite is true in the presence of the other model factors. Here the model with the congested40 dummy has a marginally higher goodness of fit level, and again the t-statistic is higher than that of the congested45 estimator.

Table 78: Incorporating Congestion Dummies in Models

	AM Peak – Model 5	PM Peak – Model 5a	PM Peak – Model 5b
Intercept	-2.710*** (t = -69.568)	-3.247*** (t = -80.849)	-3.464*** (t = -85.773)
$avgSpeed^2$	-0.0005*** (t = -144.685)	-0.0003*** (t = -80.352)	-0.0002*** (t = -49.912)
tollAmount	-0.436*** (t = -279.080)	-0.356*** (t = -188.594)	-0.362*** (t = -191.343)
transponderCount	-0.0003*** (t = -17.069)	0.003*** (t = 131.399)	0.003*** (t = 133.269)
HOT: congested50	2.054*** (t = 273.573)		
HOT: congested45		1.386*** (t = 232.182)	
HOT: congested40			1.374*** (t = 252.782)
HOT: log(hhIncomedollars)	0.137*** (t = 35.359)	0.123*** (t = 30.918)	0.124*** (t = 31.049)
HOT: hhEdu	-0.117*** (t = -39.654)	-0.187*** (t = -61.025)	-0.195*** (t = -63.643)
HOT: hhAge	-0.030*** (t = -16.084)	-0.006*** (t = -3.268)	-0.006*** (t = -3.256)
HOT: hhSize	-0.043*** (t = -34.549)	-0.021*** (t = -16.405)	-0.021*** (t = -16.263)
HOT: segmentCount	0.692*** (t = 292.457)	0.948*** (t = 394.284)	0.998*** (t = 405.242)
HOT Share	0.5287	0.5418	0.5418
Observations	1,105,171	1,191,877	1,191,877
R2	0.121	0.199	0.206
Log Likelihood	-671,423.80	-658,137.90	-652,655.80

*p<0.1; **p<0.05; ***p<0.01

Model 6 –Month of Year Dummy Variables

With January as the reference category, the next pair of models used dummy variables to indicate the month of the year in which the trip took place. Table 79 presents the results of these model estimates. In both the AM and PM peak models, the goodness of fit improves very slightly (by 0.001 in the AM and 0.005 in the PM). Each of the dummy variables is significant at the 95% confidence level, as are the rest of the factors in the models.

Table 79: Adding Month Dummy Variables

	AM Peak – Model 6a	PM Peak – Model 6a
Intercept	-2.812*** (t = -71.173)	-3.447*** (t = -84.015)
$avgSpeed^2$	-0.0005*** (t = -143.724)	-0.0002*** (t = -50.101)
tollAmount	-0.437*** (t = -279.088)	-0.423*** (t = -202.483)
transponderCount	-0.0003*** (t = -17.819)	0.003*** (t = 143.899)
HOT: congested50	2.043*** (t = 271.004)	
HOT: congested40		1.368*** (t = 249.511)
HOT: log(hhIncomedollars)	0.137*** (t = 35.333)	0.123*** (t = 30.722)
HOT: hhEdu	-0.117*** (t = -39.639)	-0.198*** (t = -64.367)
HOT: hhAge	-0.030*** (t = -16.181)	-0.006*** (t = -3.368)
HOT: hhSize	-0.044*** (t = -34.568)	-0.021*** (t = -16.317)
HOT: segmentCount	0.694*** (t = 292.088)	1.056*** (t = 406.739)
HOT: february	0.070*** (t = 6.903)	-0.059*** (t = -5.576)
HOT: march	0.077*** (t = 7.818)	-0.043*** (t = -4.141)
HOT: april	0.160*** (t = 16.008)	-0.257*** (t = -24.349)
HOT: may	0.142*** (t = 14.391)	-0.358*** (t = -34.483)
HOT: june	0.096*** (t = 9.240)	-0.350*** (t = -32.699)
HOT: july	0.091*** (t = 9.146)	-0.328*** (t = -31.108)
HOT: august	0.149*** (t = 14.938)	-0.164*** (t = -15.748)
HOT: september	0.189*** (t = 18.588)	0.188*** (t = 17.635)
HOT: october	0.212*** (t = 20.949)	0.226*** (t = 21.341)
HOT: november	0.123*** (t = 11.959)	0.157*** (t = 14.663)
HOT: december	0.020* (t = 1.898)	0.137*** (t = 12.785)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R^2	0.122	0.211
Log Likelihood	-670,994.30	-648,711.70

*p<0.1; **p<0.05; ***p<0.01

For the monthly dummy indicators in the morning peak, the odds ratio measures ranged from 1.02 (December) to 1.24 (October). In the afternoon peak, the minimum value was 0.699 (May) while the maximum was 1.25 (October).

Model 6b – Seasonal instead of Monthly Dummy Variables

Table 80 presents the model estimate results with seasonal rather than monthly dummy variables. Here 'spring' refers to March, April, and May, 'summer' refers to June, July, and August, 'fall' refers to September, October, and November, and 'winter' refers to December, January and February. Model goodness of fit levels decrease by 0.001 in both the morning and afternoon cases relative to the previous Model 6a estimates; the change in variables has an irrelevant impact on the pseudo-R^2 values. Trips in the spring, summer, or fall see an increase in

probability of Express Lane use in the morning, relative to the winter season trips. In the afternoon peak, only fall trips have a higher HOT probability than winter trips; spring and summer probabilities are lower.

Table 80: Adding Season dummy variables

	AM Peak – Model 6b	PM Peak – Model 6b
Intercept	-2.782*** (t = -71.171)	-3.412*** (t = -84.021)
$avgSpeed^2$	-0.0005*** (t = -144.499)	-0.0002*** (t = -49.383)
tollAmount	-0.437*** (t = -279.398)	-0.409*** (t = -201.872)
transponderCount	-0.0003*** (t = -16.830)	0.003*** (t = 144.118)
HOT: congested50	2.043*** (t = 271.200)	
HOT: congested40		1.369*** (t = 250.842)
HOT: log(hhIncomedollars)	0.137*** (t = 35.307)	0.124*** (t = 30.855)
HOT: hhEdu	-0.117*** (t = -39.660)	-0.198*** (t = -64.378)
HOT: hhAge	-0.030*** (t = -16.170)	-0.006*** (t = -3.342)
HOT: hhSize	-0.043*** (t = -34.553)	-0.021*** (t = -16.262)
HOT: segmentCount	0.694*** (t = 292.592)	1.043*** (t = 407.222)
HOT: spring	0.097*** (t = 16.501)	-0.243*** (t = -39.271)
HOT: summer	0.083*** (t = 13.973)	-0.300*** (t = -48.270)
HOT: fall	0.146*** (t = 24.311)	0.150*** (t = 23.892)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R2	0.122	0.21
Log Likelihood	-671,117.00	-649,541.10

*p<0.1; **p<0.05; ***p<0.01

Model 7 – Day of Week Dummy Variables

Model 7 builds upon the previous iterations by adding dummy variables for the day of the week, with Monday as the base alternative. Table 81 illustrates the improvement in model fit for both the AM and PM peak models. Again, all of the additional variables achieve significance at the 95% confidence level in both segments. In the morning peak, Tuesday, Wednesday, and Thursday trips result in a higher probability of HOT use than Monday trips, while Friday trips see the opposite effect. For afternoon trips, both Thursday and Friday trips have lower HOT use probabilities than Monday trips.

Table 81: Adding Day of Week dummy variables

	AM Peak – Model 7	PM Peak – Model 7
Intercept	-2.840*** (t = -71.090)	-3.435*** (t = -83.252)
avgSpeed^2	-0.0005*** (t = -143.044)	-0.0002*** (t = -49.831)
tollAmount	-0.480*** (t = -291.752)	-0.418*** (t = -198.727)
transponderCount	-0.0004*** (t = -24.746)	0.003*** (t = 141.262)
HOT: congested50	2.054*** (t = 271.283)	
HOT: congested40		1.365*** (t = 248.432)
HOT: log(hhIncomedollars)	0.137*** (t = 35.351)	0.124*** (t = 30.889)
HOT: hhEdu	-0.116*** (t = -38.815)	-0.199*** (t = -64.494)
HOT: hhAge	-0.033*** (t = -17.365)	-0.006*** (t = -3.298)
HOT: hhSize	-0.043*** (t = -34.038)	-0.021*** (t = -16.370)
HOT: segmentCount	0.723*** (t = 299.101)	1.056*** (t = 406.578)
HOT: february	0.106*** (t = 10.301)	-0.062*** (t = -5.870)
HOT: march	0.121*** (t = 12.146)	-0.049*** (t = -4.636)
HOT: april	0.207*** (t = 20.455)	-0.262*** (t = -24.772)
HOT: may	0.190*** (t = 19.087)	-0.360*** (t = -34.624)
HOT: june	0.122*** (t = 11.674)	-0.353*** (t = -32.912)
HOT: july	0.099*** (t = 9.854)	-0.339*** (t = -32.005)
HOT: august	0.187*** (t = 18.568)	-0.169*** (t = -16.193)
HOT: september	0.256*** (t = 25.013)	0.185*** (t = 17.316)
HOT: october	0.252*** (t = 24.656)	0.220*** (t = 20.785)
HOT: november	0.183*** (t = 17.609)	0.157*** (t = 14.590)
HOT: december	0.058*** (t = 5.570)	0.137*** (t = 12.774)
HOT: tuesday	0.128*** (t = 19.476)	0.023*** (t = 3.427)
HOT: wednesday	0.162*** (t = 24.293)	0.033*** (t = 4.816)
HOT: thursday	0.156*** (t = 23.479)	-0.089*** (t = -12.633)
HOT: friday	-0.477*** (t = -67.623)	-0.065*** (t = -8.748)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R^2	0.129	0.211
Log Likelihood	-665,422.20	-648,480.70

*p<0.1; **p<0.05; ***p<0.01

Morning peak period trips see the largest change in Express Lane use probability (relative to Mondays) on Fridays; the odds ratio value for that indicator is 0.621. In the afternoon peak, the largest decrease occurs on Friday as well (odds ratio of 0.937).

Model 8 – Hour of Day Dummy Variables

Table 82 presents the results of models estimated with dummy variables indicating the hour in which the trip was taken. In both cases, the first hour of the peak period (6am and 3pm) was selected as the base alternative. Goodness of fit values improve once again, though not all of the additional variables achieve significance at the 95% confidence level. For the morning peak

trips, both seven and eight AM trips have a higher probability of HOT use, while nine AM trips do not yield a significant impact. Afternoon trips result in higher HOT choice probabilities in the five and six PM hours.

Table 82: Adding Hour of Day Dummy Variables

	AM Peak – Model 8	PM Peak – Model 8
Intercept	-3.337*** (t = -82.311)	-3.613*** (t = -85.961)
$avgSpeed^2$	-0.0004*** (t = -112.800)	-0.0002*** (t = -47.996)
tollAmount	-0.550*** (t = -301.302)	-0.452*** (t = -196.586)
transponderCount	-0.001*** (t = -32.063)	0.004*** (t = 142.939)
HOT: congested50	1.837*** (t = 236.965)	
HOT: congested40		1.370*** (t = 248.036)
HOT: log(hhIncomedollars)	0.139*** (t = 35.360)	0.125*** (t = 31.130)
HOT: hhEdu	-0.124*** (t = -41.284)	-0.208*** (t = -67.359)
HOT: hhAge	-0.030*** (t = -16.063)	-0.006*** (t = -2.955)
HOT: hhSize	-0.045*** (t = -34.934)	-0.020*** (t = -15.264)
HOT: segmentCount	0.809*** (t = 316.730)	1.092*** (t = 407.630)
HOT: february	0.171*** (t = 16.532)	-0.071*** (t = -6.638)
HOT: march	0.159*** (t = 15.770)	-0.057*** (t = -5.430)
HOT: april	0.266*** (t = 26.052)	-0.274*** (t = -25.749)
HOT: may	0.263*** (t = 26.141)	-0.373*** (t = -35.713)
HOT: june	0.182*** (t = 17.285)	-0.370*** (t = -34.359)
HOT: july	0.109*** (t = 10.744)	-0.359*** (t = -33.753)
HOT: august	0.280*** (t = 27.502)	-0.157*** (t = -14.936)
HOT: september	0.388*** (t = 37.390)	0.234*** (t = 21.738)
HOT: october	0.380*** (t = 36.838)	0.273*** (t = 25.552)
HOT: november	0.279*** (t = 26.619)	0.215*** (t = 19.824)
HOT: december	0.119*** (t = 11.364)	0.196*** (t = 18.131)
HOT: tuesday	0.177*** (t = 26.555)	0.017*** (t = 2.585)
HOT: wednesday	0.194*** (t = 28.767)	0.028*** (t = 4.082)
HOT: thursday	0.188*** (t = 28.044)	-0.108*** (t = -15.119)
HOT: friday	-0.630*** (t = -86.912)	-0.085*** (t = -11.255)
HOT: sevenAm	0.782*** (t = 120.569)	
HOT: eightAm	0.630*** (t = 96.211)	
HOT: nineAm	0.028*** (t = 4.014)	
HOT: fourPm		-0.046*** (t = -6.531)
HOT: fivePm		0.162*** (t = 21.959)
HOT: sixPm		0.399*** (t = 57.842)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R2	0.141	0.215
Log Likelihood	-656,487.40	-645,592.70

*p<0.1; **p<0.05; ***p<0.01

In the morning peak period, the largest odds ratio value occurs with the sevenAm indicator (2.19). In the afternoon peak, the highest value occurs with the sixPm indicator (1.49).

Model 9 – Half-Hour Dummy Variables

Model 9 modifies Model 8 by including half-hour rather than hour-long dummy variables. Table 83 illustrates the pseudo-R^2 improvement that this change yields. The coefficients all achieve significance at the 95% confidence level in the morning: the probability of taking a toll lane trip increases with any start time interval besides 6:00-6:29 AM. Toll lane trip probability relative to a 3:00-3:29 PM start increases in the afternoon after 5:00 PM.

Table 83: Half-Hour Dummies instead of Hour Dummies

	AM Peak – Model 9	PM Peak – Model 9
Intercept	-4.293*** (t = -103.290)	-3.603*** (t = -84.686)
$avgSpeed^2$	-0.0004*** (t = -115.415)	-0.0002*** (t = -46.776)
tollAmount	-0.677*** (t = -331.197)	-0.458*** (t = -194.910)
transponderCount	-0.001*** (t = -41.514)	0.004*** (t = 139.094)
HOT: congested50	1.599*** (t = 203.142)	
HOT: congested40		1.367*** (t = 247.301)
HOT: log(hhIncomedollars)	0.142*** (t = 35.689)	0.125*** (t = 31.171)
HOT: hhEdu	-0.131*** (t = -42.894)	-0.209*** (t = -67.619)
HOT: hhAge	-0.028*** (t = -14.334)	-0.005*** (t = -2.860)
HOT: hhSize	-0.043*** (t = -33.016)	-0.020*** (t = -15.289)
HOT: segmentCount	0.928*** (t = 339.761)	1.098*** (t = 407.400)
HOT: february	0.265*** (t = 25.120)	-0.071*** (t = -6.659)
HOT: march	0.235*** (t = 22.909)	-0.057*** (t = -5.423)
HOT: april	0.380*** (t = 36.536)	-0.275*** (t = -25.829)
HOT: may	0.379*** (t = 37.031)	-0.374*** (t = -35.798)
HOT: june	0.274*** (t = 25.603)	-0.372*** (t = -34.479)
HOT: july	0.149*** (t = 14.406)	-0.360*** (t = -33.741)
HOT: august	0.414*** (t = 39.876)	-0.153*** (t = -14.575)
HOT: september	0.586*** (t = 55.155)	0.243*** (t = 22.492)
HOT: october	0.584*** (t = 55.337)	0.283*** (t = 26.387)
HOT: november	0.452*** (t = 42.144)	0.224*** (t = 20.639)
HOT: december	0.217*** (t = 20.271)	0.205*** (t = 18.940)
HOT: tuesday	0.237*** (t = 35.067)	0.017** (t = 2.509)
HOT: wednesday	0.242*** (t = 35.404)	0.028*** (t = 4.093)
HOT: thursday	0.235*** (t = 34.520)	-0.108*** (t = -15.107)
HOT: friday	-0.885*** (t = -116.678)	-0.086*** (t = -11.291)
HOT:am630	1.589*** (t = 165.693)	
HOT: am700	1.933*** (t = 187.112)	
HOT: am730	1.988*** (t = 188.155)	
HOT:am800	1.840*** (t = 176.059)	
HOT:am830	1.593*** (t = 155.430)	
HOT:am900	1.122*** (t = 112.116)	
HOT:am930	0.363*** (t = 35.093)	
HOT:pm1530		-0.055*** (t = -5.502)
HOT:pm1600		-0.136*** (t = -13.554)
HOT:pm1630		-0.014 (t = -1.337)
HOT:pm1700		0.077*** (t = 7.469)
HOT:pm1730		0.207*** (t = 20.287)
HOT:pm1800		0.405*** (t = 40.743)
HOT:pm1830		0.345*** (t = 35.451)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R2	0.162	0.215
Log Likelihood	-640,584.30	-645,316.00

*p<0.1; **p<0.05; ***p<0.01

Within the morning peak period, the largest odds ratios occur among the am700, am730, and am800 indicators: all three of those values exceed an odds ratio of six. In the afternoon peak, the largest odds ratio is found with the pm1800 indicator: the value there is 1.50.

Model 10 – Square of Toll Amount

The motivation behind this pair of models was to examine whether a user's response to the Express Lane toll may be non-linear, as was the case for the average speed difference variable. Table 84 presents the results of the estimated models. In both cases the model fit suffers relative to Model 9; similarly, the $tollAmount^2$ coefficients have lower t-statistics than their unsquared counterparts.

Table 84: Toll Amount Squared

	AM Peak – Model 10	PM Peak – Model 10
Intercept	-4.440*** (t = -108.360)	-3.281*** (t = -77.818)
$avgSpeed^2$	-0.0004*** (t = -110.085)	-0.0002*** (t = -43.613)
$tollAmount^2$	-0.066*** (t = -287.120)	-0.047*** (t = -151.550)
transponderCount	-0.001*** (t = -59.567)	0.004*** (t = 156.218)
HOT: congested50	1.344*** (t = 178.081)	
HOT: congested40		1.231*** (t = 229.964)
HOT: log(hhIncomedollars)	0.134*** (t = 34.217)	0.124*** (t = 31.129)
HOT: hhEdu	-0.128*** (t = -42.857)	-0.205*** (t = -66.629)
HOT: hhAge	-0.027*** (t = -14.059)	-0.005** (t = -2.383)
HOT: hhSize	-0.043*** (t = -34.058)	-0.020*** (t = -15.513)
HOT: segmentCount	0.759*** (t = 306.466)	0.954*** (t = 400.957)
HOT: february	0.225*** (t = 21.793)	-0.087*** (t = -8.164)
HOT: march	0.188*** (t = 18.712)	-0.108*** (t = -10.292)
HOT: april	0.293*** (t = 28.761)	-0.306*** (t = -29.032)
HOT: may	0.311*** (t = 31.082)	-0.391*** (t = -37.647)
HOT: june	0.236*** (t = 22.468)	-0.375*** (t = -35.060)
HOT: july	0.119*** (t = 11.804)	-0.373*** (t = -35.235)
HOT: august	0.337*** (t = 33.170)	-0.270*** (t = -25.995)
HOT: september	0.522*** (t = 50.102)	0.047*** (t = 4.422)
HOT: october	0.517*** (t = 49.906)	0.074*** (t = 7.011)
HOT: november	0.400*** (t = 38.005)	0.045*** (t = 4.218)
HOT: december	0.169*** (t = 16.074)	0.033*** (t = 3.119)
HOT: tuesday	0.242*** (t = 36.444)	0.001 (t = 0.168)
HOT: wednesday	0.249*** (t = 37.069)	-0.004 (t = -0.601)
HOT: thursday	0.242*** (t = 36.193)	-0.182*** (t = -25.680)
HOT: friday	-0.619*** (t = -85.400)	-0.169*** (t = -22.424)
HOT:am630	1.242*** (t = 136.159)	
HOT: am700	1.583*** (t = 160.303)	
HOT: am730	1.633*** (t = 161.645)	
HOT:am800	1.511*** (t = 150.849)	
HOT:am830	1.351*** (t = 136.286)	
HOT:am900	1.121*** (t = 114.943)	
HOT:am930	0.790*** (t = 79.378)	
HOT:pm1530		-0.142*** (t = -14.472)
HOT:pm1600		-0.305*** (t = -30.933)
HOT:pm1630		-0.266*** (t = -26.669)
HOT:pm1700		-0.213*** (t = -21.374)
HOT:pm1730		-0.099*** (t = -10.124)
HOT:pm1800		0.123*** (t = 12.876)
HOT:pm1830		0.148*** (t = 15.571)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R^2	0.136	0.205
Log Likelihood	-660,075.60	-653,551.30

*p<0.1; **p<0.05; ***p<0.01

With the square of the toll amount factor, the corresponding odds ratio values were much closer to one than in previous models. Model 9 saw toll amount odds ratios of 0.508 (morning peak) and 0.632 (afternoon peak); Model 10 yielded values of 0.936 (morning peak) and 0.954

(afternoon peak). Note that the values of the factors themselves are squared in Model 10, so that a maximum toll amount of $8.50 in Model 9 is represented as $72.25 in Model 10.

Model 11 – htDensity instead of Transponder Counts

The last of this series of models replaces the transponderCount variable (an alternative-specific variable with a generic coefficient) with the htDensity variable. The htDensity variable divides the transponder count in the Express Lane by the length of that Express Lane segment. The variable differs from the traffic engineering definition of density as vehicles are counted over a fifteen-minute interval rather than instantaneously. Table 85 presents the results of this substitution. The morning and afternoon coefficients differ in their signs: the estimator is negative in the morning and positive in the afternoon. An increase of fifteen minute transponder density reduces the probability of using the Express Lanes in the morning, while increasing that probability in the afternoon. Both achieve significance at the 95% confidence level. The pseudo-R^2 model fit measure is higher than that of Model 9 for the AM period trips, though it is lower for the PM peak trips. Similarly, the t-statistic for the AM peak coefficient has increased relative to Model 9, while the PM peak coefficient t-statistic has decreased.

Table 85: htDensity instead of Transponder Counts

	AM Peak – Model 11	PM Peak – Model 11
Intercept	-3.608*** (t = -85.398)	-4.859*** (t = -113.439)
$avgSpeed^2$	-0.0004*** (t = -125.068)	-0.0002*** (t = -49.082)
tollAmount	-0.718*** (t = -343.980)	-0.488*** (t = -209.743)
HOT: htDensity	-0.004*** (t = -83.858)	0.008*** (t = 102.099)
HOT: congested50	1.633*** (t = 225.371)	
HOT: congested40		1.493*** (t = 278.733)
HOT: log(hhIncomedollars)	0.133*** (t = 33.323)	0.134*** (t = 33.546)
HOT: hhEdu	-0.128*** (t = -41.872)	-0.213*** (t = -69.249)
HOT: hhAge	-0.028*** (t = -14.385)	-0.007*** (t = -3.798)
HOT: hhSize	-0.043*** (t = -33.187)	-0.020*** (t = -15.381)
HOT: segmentCount	0.878*** (t = 315.853)	1.159*** (t = 396.752)
HOT: february	0.262*** (t = 24.795)	-0.038*** (t = -3.594)
HOT: march	0.246*** (t = 23.971)	0.048*** (t = 4.627)
HOT: april	0.395*** (t = 37.934)	-0.260*** (t = -24.639)
HOT: may	0.395*** (t = 38.484)	-0.352*** (t = -33.935)
HOT: june	0.280*** (t = 26.137)	-0.331*** (t = -30.906)
HOT: july	0.144*** (t = 13.956)	-0.313*** (t = -29.698)
HOT: august	0.432*** (t = 41.538)	-0.176*** (t = -16.776)
HOT: september	0.624*** (t = 58.482)	0.119*** (t = 10.977)
HOT: october	0.625*** (t = 58.913)	0.132*** (t = 12.263)
HOT: november	0.501*** (t = 46.548)	0.022** (t = 2.029)
HOT: december	0.238*** (t = 22.203)	0.012 (t = 1.088)
HOT: tuesday	0.231*** (t = 34.244)	0.067*** (t = 10.059)
HOT: wednesday	0.231*** (t = 33.948)	0.102*** (t = 15.091)
HOT: thursday	0.221*** (t = 32.609)	0.120*** (t = 17.445)
HOT: friday	-0.949*** (t = -124.485)	0.261*** (t = 36.633)
HOT:am630	1.688*** (t = 173.924)	
HOT: am700	2.052*** (t = 196.051)	
HOT: am730	2.096*** (t = 196.150)	
HOT:am800	1.904*** (t = 182.158)	
HOT:am830	1.607*** (t = 158.432)	
HOT:am900	1.080*** (t = 109.233)	
HOT:am930	0.269*** (t = 26.182)	
HOT:pm1530		0.144*** (t = 14.715)
HOT:pm1600		0.148*** (t = 15.371)
HOT:pm1630		0.224*** (t = 22.548)
HOT:pm1700		0.280*** (t = 27.640)
HOT:pm1730		0.343*** (t = 33.706)
HOT:pm1800		0.479*** (t = 48.100)
HOT:pm1830		0.365*** (t = 37.320)
HOT Share	0.5287	0.5418
Observations	1,105,171	1,191,877
R^2	0.166	0.209
Log Likelihood	-637,634.60	-649,812.10

*p<0.1; **p<0.05; ***p<0.01

Replacing the difference in transponder count factor with the Express Lane fifteen minute transponder density variable yields very little change in the odds ratios, though the two measures

are not strictly comparable. In both cases, the odds ratios in the morning and afternoon peak are indistinguishable from a value of one.

Overview of Models with Additional Variables

The iterative building of the models in this section generally came with model fit improvements as more variables were added. For this reason, the Akaike Information Criterion measure was also examined to measure the benefit of the additional variable load in the models. This measure improved understanding of the differences between the models. For example, Model 6a, which included monthly dummy variables, had eight additional variables relative to Model 6b, which included the seasonal dummy variables. The R^2 improvement in Model 6a was very small: 0.001 in both cases. The AIC measure, which was lower for Model 6a in both peak periods, reinforced the benefit of the additional monthly variables rather than the smaller set of seasonal variables. Similarly, Model 9 used half-hour time increments to look at the impact of trip start times, while Model 8 used hour-long time increments. In both cases, the additional variables improved the model fit; the AIC measure further suggested that the benefits of these variables outweighed the cost. Table 86 presents all of the models with their respective R^2 and AIC measures. Models 4a, 4b, and 4c were not included in the AIC comparison as they were all univariate models meant to compare the effects of those individual factors. The highlighted values show the models with the highest pseudo-R^2 value and the lowest AIC values.

Table 86: Summary of Models with Additional Variables

Model	AM Peak - R^2 Value	AM Peak – AIC	PM Peak – R^2 Value	PM Peak – AIC
Model 2	0.102	1,461,537	0.205	1,310,389
Model 3	0.106	1,455,090	0.206	1,309,062
Model 4a	0.028	N/A	0.074	N/A
Model 4b	0.027	N/A	0.068	N/A
Model 4c	0.021	N/A	0.056	N/A
Model 5a	0.121	1,435,254	0.199	1,320,109
Model 5b		N/A	0.206	1,309,062
Model 6a	0.122	1,434,016	0.211	1,301,199
Model 6b	0.122	1,434,280	0.210	1,302,829
Model 7	0.129	1,421,294	0.211	1,300,734
Model 8	0.141	1,399,401	0.215	1,294,921
Model 9	0.162	1,364,041	0.215	1,294,371
Model 10	0.136	1,406,813	0.205	1,310,834
Model 11	0.166	1,358,037	0.209	1,303,683

Though Model 11, in which the transponderCount variable is replaced by the htDensity variable, yields the highest R^2 and lowest AIC values in the AM peak, the author selected Model 9 as the base model going forward. Though the model with htDensity performed better in the morning, the exclusion of the transponderCount variable means that the model has no component representing vehicle counts in the General Purpose lanes. The author felt that this shortcoming was not worth the 0.004 increase in R^2. The remainder of the models in this chapter are built off of Model 9 for both peak periods for this reason.

Interaction Terms

After adding additional variables that were not included in the Sheikh, 2015 TRB paper, the next step in the modeling investigation was to examine interaction terms that had not previously been investigated. As stated above, Model 9 served as the base for both the AM and PM peak periods due to its performance and the inclusion of the transponderCount factor for both lane types.

Model 12 – Toll over log(Income)

The first of these interaction terms divides the Express Lane toll amount by the log of the household income of the user. This tollLogIncome variable served to investigate whether users

consider their own income level along with the toll amount when making lane choice decisions. Table 87 and Table 88 present the results from these models. Table 87 shows the AM peak models: 12a excludes the log of the household income on its own, while 12b includes that factor as well. In both cases the tollLogIncome coefficient is significant at the 95% confidence level. Note that the resulting odds ratios of the estimated coefficients are very small: less than 0.01 in both Models 12a and 12b, indicating a large decrease in HOT use probability given a unit increase in toll over log(household income).

Table 87: Toll Over log(Income) – AM Peak Models

	AM Peak – Model 12a	AM Peak – Model 12b
Intercept	-2.588*** (t = -124.014)	-2.007*** (t = -48.434)
$avgSpeed^2$	-0.0004*** (t = -114.661)	-0.0004*** (t = -114.651)
tollLogIncome	-7.343*** (t = -329.696)	-7.390*** (t = -328.817)
transponderCount	-0.001*** (t = -42.740)	-0.001*** (t = -42.129)
HOT: congested50	1.589*** (t = 202.287)	1.591*** (t = 202.530)
HOT: log(hhIncomedollars)	()	-0.065*** (t = -16.231)
HOT: hhEdu	-0.144*** (t = -49.921)	-0.128*** (t = -41.991)
HOT: hhAge	-0.033*** (t = -17.776)	-0.027*** (t = -14.172)
HOT: hhSize	-0.049*** (t = -41.034)	-0.041*** (t = -31.860)
HOT: segmentCount	0.917*** (t = 339.795)	0.923*** (t = 338.657)
HOT: february	0.263*** (t = 24.999)	0.263*** (t = 25.000)
HOT: march	0.234*** (t = 22.857)	0.234*** (t = 22.828)
HOT: april	0.376*** (t = 36.245)	0.377*** (t = 36.285)
HOT: may	0.374*** (t = 36.575)	0.375*** (t = 36.650)
HOT: june	0.269*** (t = 25.131)	0.269*** (t = 25.158)
HOT: july	0.145*** (t = 14.062)	0.145*** (t = 14.064)
HOT: august	0.405*** (t = 39.077)	0.407*** (t = 39.242)
HOT: september	0.574*** (t = 54.102)	0.577*** (t = 54.347)
HOT: october	0.572*** (t = 54.244)	0.575*** (t = 54.484)
HOT: november	0.440*** (t = 41.102)	0.443*** (t = 41.334)
HOT: december	0.206*** (t = 19.279)	0.208*** (t = 19.487)
HOT: tuesday	0.237*** (t = 35.128)	0.237*** (t = 35.070)
HOT: wednesday	0.243*** (t = 35.614)	0.242*** (t = 35.511)
HOT: thursday	0.236*** (t = 34.686)	0.235*** (t = 34.578)
HOT: friday	-0.865*** (t = -114.512)	-0.871*** (t = -115.163)
HOT:am630	1.561*** (t = 163.527)	1.569*** (t = 164.059)
HOT: am700	1.902*** (t = 185.079)	1.912*** (t = 185.617)
HOT: am730	1.958*** (t = 186.199)	1.967*** (t = 186.714)
HOT:am800	1.816*** (t = 174.362)	1.823*** (t = 174.792)
HOT:am830	1.576*** (t = 154.242)	1.580*** (t = 154.545)
HOT:am900	1.117*** (t = 111.882)	1.118*** (t = 111.961)
HOT:am930	0.371*** (t = 35.984)	0.369*** (t = 35.749)
HOT Share	0.5287	0.5287
Observations	1,105,171	1,105,171
R^2	0.16	0.16
Log Likelihood	-641,814.40	-641,682.60

*p<0.1; **p<0.05; ***p<0.01

Table 88 presents the PM peak results. Again, the tollLogIncome coefficients are significant at the 95% confidence level both with and without the presence of the separate household income variable. For both the AM and PM peak periods, the goodness of fit measures decrease by 0.001 with the inclusion of the tollLogIncome factor; this occurs with or without the

inclusion of household income on its own. The AIC measures for all four Model 12 variants also exceed those of Model 9. Again, the odds ratios for this factor are less than 0.01 in both cases.

Table 88: Toll Over log(Income) – PM Peak Models

	PM Peak – Model 12a	PM Peak – Model 12b
Intercept	-2.296*** (t = -105.967)	-2.640*** (t = -62.804)
$avgSpeed^2$	-0.0002*** (t = -47.050)	-0.0002*** (t = -46.871)
tollLogIncome	-5.047*** (t = -195.387)	-5.023*** (t = -193.582)
transponderCount	0.004*** (t = 139.612)	0.004*** (t = 139.813)
HOT: congested40	1.362*** (t = 246.788)	1.362*** (t = 246.763)
HOT: log(hhIncomedollars)		0.039*** (t = 9.558)
HOT: hhEdu	-0.198*** (t = -67.718)	-0.207*** (t = -67.076)
HOT: hhAge	-0.001 (t = -0.597)	-0.005** (t = -2.452)
HOT: hhSize	-0.015*** (t = -11.987)	-0.019*** (t = -14.690)
HOT: segmentCount	1.098*** (t = 409.382)	1.096*** (t = 406.907)
HOT: february	-0.071*** (t = -6.681)	-0.072*** (t = -6.699)
HOT: march	-0.058*** (t = -5.491)	-0.059*** (t = -5.532)
HOT: april	-0.276*** (t = -25.948)	-0.275*** (t = -25.908)
HOT: may	-0.375*** (t = -35.880)	-0.375*** (t = -35.854)
HOT: june	-0.373*** (t = -34.625)	-0.373*** (t = -34.602)
HOT: july	-0.361*** (t = -33.887)	-0.361*** (t = -33.900)
HOT: august	-0.156*** (t = -14.831)	-0.158*** (t = -14.977)
HOT: september	0.239*** (t = 22.171)	0.237*** (t = 21.914)
HOT: october	0.279*** (t = 26.043)	0.276*** (t = 25.782)
HOT: november	0.220*** (t = 20.321)	0.218*** (t = 20.101)
HOT: december	0.202*** (t = 18.616)	0.200*** (t = 18.423)
HOT: tuesday	0.017** (t = 2.484)	0.016** (t = 2.457)
HOT: wednesday	0.027*** (t = 4.036)	0.027*** (t = 3.963)
HOT: thursday	-0.109*** (t = -15.221)	-0.110*** (t = -15.415)
HOT: friday	-0.087*** (t = -11.458)	-0.089*** (t = -11.658)
HOT:pm1530	-0.058*** (t = -5.801)	-0.058*** (t = -5.849)
HOT:pm1600	-0.141*** (t = -14.004)	-0.142*** (t = -14.144)
HOT:pm1630	-0.019* (t = -1.897)	-0.022** (t = -2.110)
HOT:pm1700	0.071*** (t = 6.867)	0.068*** (t = 6.596)
HOT:pm1730	0.201*** (t = 19.649)	0.198*** (t = 19.344)
HOT:pm1800	0.398*** (t = 40.136)	0.396*** (t = 39.862)
HOT:pm1830	0.340*** (t = 34.971)	0.339*** (t = 34.816)
HOT Share	0.5418	0.5418
Observations	1,191,877	1,191,877
R^2	0.215	0.215
Log Likelihood	-645,630.40	-645,584.80

*p<0.1; **p<0.05; ***p<0.01

Model 13 – Toll over Income

Table 89 and Table 90 present models similar to the previous set, but with toll divided by the unmodified household income rather than the log of household income. The resulting models have poorer goodness of fit measures in all four cases. The AIC measures increase for both the

AM and PM peak periods as well. The odds ratio measures are zero in both cases, due to the very small magnitude of toll/income values (a unit increase is very unlikely).

Table 89: Toll over Income - AM Peak Models

	AM Peak – Model 13a	AM Peak – Model 13b
Intercept	-1.736*** (t = -87.391)	0.325*** (t = 6.228)
avgSpeed^2	-0.0002*** (t = -70.283)	-0.0002*** (t = -71.022)
tollIncome	-2,671.174*** (t = -95.458)	-3,721.864*** (t = -96.688)
transponderCount	-0.001*** (t = -67.629)	-0.001*** (t = -66.211)
HOT: congested50	0.993*** (t = 139.310)	1.010*** (t = 141.313)
HOT: log(hhIncomedollars)		-0.215*** (t = -42.622)
HOT: hhEdu	-0.141*** (t = -50.759)	-0.107*** (t = -36.887)
HOT: hhAge	-0.029*** (t = -16.329)	-0.016*** (t = -9.055)
HOT: hhSize	-0.048*** (t = -41.853)	-0.029*** (t = -23.866)
HOT: segmentCount	0.441*** (t = 219.774)	0.462*** (t = 222.802)
HOT: february	0.162*** (t = 16.203)	0.164*** (t = 16.412)
HOT: march	0.145*** (t = 14.981)	0.146*** (t = 15.117)
HOT: april	0.167*** (t = 16.970)	0.172*** (t = 17.484)
HOT: may	0.162*** (t = 16.754)	0.166*** (t = 17.207)
HOT: june	0.164*** (t = 16.238)	0.163*** (t = 16.147)
HOT: july	0.118*** (t = 12.126)	0.116*** (t = 11.877)
HOT: august	0.183*** (t = 18.739)	0.187*** (t = 19.132)
HOT: september	0.237*** (t = 23.800)	0.244*** (t = 24.468)
HOT: october	0.234*** (t = 23.575)	0.240*** (t = 24.217)
HOT: november	0.122*** (t = 12.199)	0.128*** (t = 12.772)
HOT: december	-0.075*** (t = -7.473)	-0.070*** (t = -6.943)
HOT: tuesday	0.200*** (t = 31.177)	0.200*** (t = 31.170)
HOT: wednesday	0.236*** (t = 36.330)	0.235*** (t = 36.165)
HOT: thursday	0.223*** (t = 34.572)	0.222*** (t = 34.369)
HOT: friday	-0.089*** (t = -13.398)	-0.115*** (t = -17.216)
HOT:am630	0.550*** (t = 65.593)	0.579*** (t = 68.835)
HOT: am700	0.722*** (t = 80.799)	0.760*** (t = 84.600)
HOT: am730	0.760*** (t = 83.148)	0.799*** (t = 86.907)
HOT:am800	0.807*** (t = 87.191)	0.838*** (t = 90.129)
HOT:am830	0.928*** (t = 99.352)	0.947*** (t = 101.151)
HOT:am900	0.946*** (t = 102.444)	0.950*** (t = 102.809)
HOT:am930	0.754*** (t = 80.076)	0.741*** (t = 78.557)
HOT Share	0.5287	0.5287
Observations	1,105,171	1,105,171
R^2	0.081	0.082
Log Likelihood	-702,319.60	-701,405.50

*p<0.1; **p<0.05; ***p<0.01

Table 90 presents the results from the PM peak models with the tollIncome interaction term. As was the case with the AM peak models, the tollIncome coefficients are significant at the 95% confidence level. Also notable in Model 13b is the negative sign on the household income coefficient; in Model 12b, in which the toll amount was divided by the log of the household income, this coefficient was positive. Again the odds ratio measures are zero in both cases, for the reason given above.

Table 90: Toll over Income - PM Peak Models

	PM Peak – Model 13a	PM Peak – Model 13b
Intercept	-1.332*** (t = -63.850)	-0.471*** (t = -9.180)
$avgSpeed^2$	-0.0001*** (t = -36.600)	-0.0001*** (t = -37.174)
tollIncome	-3,036.830*** (t = -70.929)	-3,653.682*** (t = -66.317)
transponderCount	0.005*** (t = 191.833)	0.005*** (t = 189.918)
HOT: congested40	1.061*** (t = 206.783)	1.067*** (t = 207.465)
HOT: log(hhIncomedollars)		-0.092*** (t = -18.373)
HOT: hhEdu	-0.192*** (t = -65.951)	-0.176*** (t = -57.970)
HOT: hhAge	-0.004** (t = -2.272)	0.002 (t = 1.031)
HOT: hhSize	-0.021*** (t = -16.970)	-0.012*** (t = -9.264)
HOT: segmentCount	0.857*** (t = 389.581)	0.866*** (t = 383.099)
HOT: february	-0.096*** (t = -9.204)	-0.096*** (t = -9.136)
HOT: march	-0.147*** (t = -14.133)	-0.144*** (t = -13.886)
HOT: april	-0.314*** (t = -30.119)	-0.314*** (t = -30.115)
HOT: may	-0.395*** (t = -38.517)	-0.395*** (t = -38.535)
HOT: june	-0.365*** (t = -34.582)	-0.367*** (t = -34.757)
HOT: july	-0.380*** (t = -36.337)	-0.381*** (t = -36.409)
HOT: august	-0.395*** (t = -38.519)	-0.390*** (t = -37.954)
HOT: september	-0.211*** (t = -20.358)	-0.200*** (t = -19.202)
HOT: october	-0.196*** (t = -19.142)	-0.184*** (t = -17.931)
HOT: november	-0.188*** (t = -17.951)	-0.178*** (t = -16.969)
HOT: december	-0.173*** (t = -16.545)	-0.164*** (t = -15.678)
HOT: tuesday	-0.019*** (t = -2.875)	-0.018*** (t = -2.707)
HOT: wednesday	-0.043*** (t = -6.398)	-0.040*** (t = -6.071)
HOT: thursday	-0.289*** (t = -41.362)	-0.283*** (t = -40.472)
HOT: friday	-0.288*** (t = -38.696)	-0.282*** (t = -37.795)
HOT:pm1530	-0.220*** (t = -22.837)	-0.217*** (t = -22.465)
HOT:pm1600	-0.451*** (t = -46.761)	-0.444*** (t = -45.919)
HOT:pm1630	-0.505*** (t = -51.980)	-0.493*** (t = -50.576)
HOT:pm1700	-0.515*** (t = -53.597)	-0.500*** (t = -51.776)
HOT:pm1730	-0.427*** (t = -45.419)	-0.411*** (t = -43.452)
HOT:pm1800	-0.156*** (t = -16.899)	-0.142*** (t = -15.338)
HOT:pm1830	0.015 (t = 1.568)	0.022** (t = 2.375)
HOT Share	0.5418	0.5418
Observations	1,191,877	1,191,877
R^2	0.193	0.194
Log Likelihood	-662,981.50	-662,812.40

*p<0.1; **p<0.05; ***p<0.01

Model 14 – Income over Household Size

The next potential interaction explored was that of household income over household size. The motivation behind this term was to investigate whether per-person income was a better determinant of lane choice decisions: a household making $100,000 annually may behave differently if it has two persons versus five persons, for example. Table 91 presents the results from four variants of this model. The first two, Models 14a and 14b, include the household income over household size interaction term. Model 14b includes household income and household size separately as well, whereas Model 14a does not. Models 14c and 14d follow this same pattern but use the log of the household income in the interaction term. In all four models, the odds ratio of the new factors are very close to one: the highest odds ratio value occurs in Model14d, in which log(household income)/household size yields an odds ratio of 1.05.

Table 91: Income over Household Size - AM Peak Models

	AM Peak – Model 14a	AM Peak – Model 14b	AM Peak – Model 14c	AM Peak – Model 14d
Intercept	-3.025*** (t = -144.869)	-1.170*** (t = -16.543)	-3.252*** (t = -149.712)	-4.541*** (t = -100.381)
$avgSpeed^2$	-0.0004*** (t = -115.748)	-0.0004*** (t = -115.499)	-0.0004*** (t = -114.715)	-0.0004*** (t = -115.451)
tollIncome	-0.679*** (t = -331.663)	-0.680*** (t = -331.910)	-0.676*** (t = -331.010)	-0.677*** (t = -331.130)
transponderCount	-0.001*** (t = -41.427)	-0.001*** (t = -40.767)	-0.001*** (t = -40.670)	-0.001*** (t = -41.539)
HOT: congested50	1.602*** (t = 203.308)	1.601*** (t = 203.147)	1.594*** (t = 202.774)	1.599*** (t = 203.173)
HOT: hhEdu	-0.149*** (t = -49.655)	-0.139*** (t = -45.341)	-0.098*** (t = -34.170)	-0.134*** (t = -43.707)
HOT: hhAge	-0.035*** (t = -18.128)	-0.031*** (t = -15.993)	-0.012*** (t = -6.411)	-0.025*** (t = -13.113)
HOT:I(hhIncomeDollars/hhSize)	0.00001*** (t = 62.512)	0.00002*** (t = 54.194)		
HOT:log(hhIncomeDollars)		-0.209*** (t = -27.591)		0.136*** (t = 34.045)
HOT: hhSize		0.051*** (t = 23.678)		-0.007** (t = -2.267)
HOT:I(log(hhIncomeDollars)/hhSize)			0.041*** (t = 26.547)	0.050*** (t = 13.926)
HOT: segmentCount	0.927*** (t = 339.543)	0.929*** (t = 339.678)	0.932*** (t = 341.496)	0.928*** (t = 339.765)
HOT: february	0.265*** (t = 25.139)	0.265*** (t = 25.079)	0.263*** (t = 24.978)	0.265*** (t = 25.098)

Table 91 Continued

HOT: march	0.236*** (t = 22.972)	0.235*** (t = 22.930)	0.233*** (t = 22.735)	0.234*** (t = 22.888)
HOT: april	0.381*** (t = 36.613)	0.381*** (t = 36.557)	0.377*** (t = 36.317)	0.380*** (t = 36.526)
HOT: may	0.380*** (t = 37.074)	0.380*** (t = 37.020)	0.377*** (t = 36.853)	0.379*** (t = 36.998)
HOT: june	0.276*** (t = 25.755)	0.277*** (t = 25.860)	0.273*** (t = 25.554)	0.274*** (t = 25.611)
HOT: july	0.150*** (t = 14.482)	0.151*** (t = 14.576)	0.148*** (t = 14.403)	0.148*** (t = 14.388)
HOT: august	0.415*** (t = 39.918)	0.416*** (t = 40.004)	0.414*** (t = 39.921)	0.413*** (t = 39.843)
HOT: september	0.587*** (t = 55.200)	0.588*** (t = 55.280)	0.586*** (t = 55.161)	0.585*** (t = 55.092)
HOT: october	0.586*** (t = 55.401)	0.587*** (t = 55.488)	0.584*** (t = 55.335)	0.584*** (t = 55.283)
HOT: november	0.454*** (t = 42.249)	0.456*** (t = 42.383)	0.452*** (t = 42.167)	0.452*** (t = 42.101)
HOT: december	0.219*** (t = 20.407)	0.221*** (t = 20.566)	0.217*** (t = 20.292)	0.217*** (t = 20.235)
HOT: tuesday	0.237*** (t = 35.077)	0.237*** (t = 34.962)	0.236*** (t = 34.879)	0.237*** (t = 35.066)
HOT: wednesday	0.242*** (t = 35.421)	0.242*** (t = 35.318)	0.240*** (t = 35.193)	0.242*** (t = 35.399)
HOT: thursday	0.235*** (t = 34.579)	0.235*** (t = 34.493)	0.233*** (t = 34.299)	0.235*** (t = 34.525)
HOT: friday	-0.886*** (t = -116.669)	-0.887*** (t = -116.761)	-0.885*** (t = -116.778)	-0.885*** (t = -116.633)
HOT:am630	1.593*** (t = 165.926)	1.596*** (t = 166.107)	1.587*** (t = 165.568)	1.589*** (t = 165.606)
HOT: am700	1.935*** (t = 187.181)	1.938*** (t = 187.319)	1.932*** (t = 187.121)	1.931*** (t = 186.989)
HOT: am730	1.990*** (t = 188.221)	1.992*** (t = 188.306)	1.986*** (t = 188.113)	1.987*** (t = 188.083)
HOT:am800	1.844*** (t = 176.245)	1.845*** (t = 176.233)	1.836*** (t = 175.775)	1.840*** (t = 176.030)
HOT:am830	1.594*** (t = 155.405)	1.594*** (t = 155.302)	1.589*** (t = 155.206)	1.592*** (t = 155.394)
HOT:am900	1.118*** (t = 111.680)	1.116*** (t = 111.371)	1.119*** (t = 111.887)	1.120*** (t = 111.947)
HOT:am930	0.359*** (t = 34.742)	0.358*** (t = 34.638)	0.362*** (t = 35.054)	0.362*** (t = 34.999)
HOT Share	0.5278	0.5278	0.5407	0.5407
Observations	1,105,171	1,105,171	1,105,171	1,105,171
R^2	0.163	0.164	0.161	0.162
Log Likelihood	-639,461.30	-639,074.20	-641,091.50	-640,487.40

*p<0.1; **p<0.05; ***p<0.01

Table 92 presents the results from the afternoon peak models; the four models follow the same pattern as those of the AM peak. Like their morning counterparts, these models all see R^2 values that are equal to or within 0.001 of the Model 9 base, and again Model 14b has the lowest AIC measure out of all of the models examined so far. Also like the morning peak models, the resulting odds ratios range from 1.00 to 1.02.

Table 92: Income over Household Size - PM Peak Models

	PM Peak – Model 14a	PM Peak – Model 14b	PM Peak – Model 14c	PM Peak – Model 14d
Intercept	-2.444*** (t = -112.967)	-2.003*** (t = -28.326)	-2.525*** (t = -112.318)	-3.690*** (t = -79.885)
$avgSpeed^2$	-0.0002*** (t = -46.855)	-0.0002*** (t = -46.638)	-0.0002*** (t = -47.152)	-0.0002*** (t = -46.761)
tollAmount	-0.458*** (t = -194.924)	-0.459*** (t = -195.054)	-0.457*** (t = -194.615)	-0.458*** (t = -194.904)
transponderCount	0.004*** (t = 139.061)	0.004*** (t = 139.097)	0.004*** (t = 139.115)	0.004*** (t = 139.093)
HOT: congested40	1.366*** (t = 247.235)	1.367*** (t = 247.279)	1.361*** (t = 246.541)	1.367*** (t = 247.299)
HOT: hhEdu	-0.211*** (t = -69.572)	-0.213*** (t = -68.894)	-0.178*** (t = -61.036)	-0.210*** (t = -67.786)
HOT: hhAge	-0.006*** (t = -3.135)	-0.007*** (t = -3.895)	0.007*** (t = 3.641)	-0.005** (t = -2.439)
HOT:I(hhIncomeDollars/hhSize)	0.00001*** (t = 39.894)	0.00001*** (t = 28.289)		
HOT:log(hhIncomeDollars)		-0.054*** (t = -7.171)		0.123*** (t = 30.439)
HOT: hhSize		0.028*** (t = 12.937)		-0.007** (t = -2.438)
HOT:I(log(hhIncomeDollars)/hhSize)			0.011*** (t = 7.030)	0.018*** (t = 4.818)
HOT: segmentCount	1.098*** (t = 407.904)	1.097*** (t = 407.270)	1.101*** (t = 408.703)	1.098*** (t = 407.418)
HOT: february	-0.071*** (t = -6.641)	-0.071*** (t = -6.659)	-0.071*** (t = -6.647)	-0.071*** (t = -6.666)
HOT: march	-0.058*** (t = -5.424)	-0.058*** (t = -5.464)	-0.058*** (t = -5.443)	-0.058*** (t = -5.424)
HOT: april	-0.275*** (t = -25.847)	-0.275*** (t = -25.817)	-0.277*** (t = -26.046)	-0.275*** (t = -25.830)
HOT: may	-0.375*** (t = -35.817)	-0.375*** (t = -35.816)	-0.376*** (t = -35.936)	-0.374*** (t = -35.808)
HOT: june	-0.371*** (t = -34.414)	-0.371*** (t = -34.355)	-0.373*** (t = -34.555)	-0.372*** (t = -34.488)
HOT: july	-0.359*** (t = -33.656)	-0.358*** (t = -33.579)	-0.359*** (t = -33.747)	-0.360*** (t = -33.750)
HOT: august	-0.152*** (t = -14.485)	-0.152*** (t = -14.446)	-0.153*** (t = -14.567)	-0.153*** (t = -14.588)
HOT: september	0.244*** (t = 22.581)	0.245*** (t = 22.630)	0.242*** (t = 22.452)	0.243*** (t = 22.474)
HOT: october	0.284*** (t = 26.500)	0.285*** (t = 26.567)	0.282*** (t = 26.307)	0.282*** (t = 26.376)

Table 92 Continued

HOT: November	0.225*** (t = 20.759)	0.226*** (t = 20.857)	0.223*** (t = 20.570)	0.224*** (t = 20.631)
HOT: december	0.206*** (t = 19.033)	0.207*** (t = 19.126)	0.204*** (t = 18.842)	0.205*** (t = 18.932)
HOT: tuesday	0.017** (t = 2.504)	0.017** (t = 2.485)	0.017** (t = 2.493)	0.017** (t = 2.501)
HOT: wednesday	0.028*** (t = 4.104)	0.028*** (t = 4.099)	0.028*** (t = 4.127)	0.028*** (t = 4.090)
HOT: thursday	-0.108*** (t = -15.110)	-0.108*** (t = -15.135)	-0.107*** (t = -14.967)	-0.108*** (t = -15.113)
HOT: friday	-0.086*** (t = -11.318)	-0.087*** (t = -11.356)	-0.085*** (t = -11.149)	-0.086*** (t = -11.298)
HOT:pm1530	-0.054*** (t = -5.428)	-0.054*** (t = -5.375)	-0.057*** (t = -5.677)	-0.055*** (t = -5.511)
HOT:pm1600	-0.135*** (t = -13.435)	-0.134*** (t = -13.342)	-0.138*** (t = -13.712)	-0.136*** (t = -13.553)
HOT:pm1630	-0.012 (t = -1.151)	-0.011 (t = -1.044)	-0.016 (t = -1.564)	-0.013 (t = -1.313)
HOT:pm1700	0.079*** (t = 7.667)	0.080*** (t = 7.777)	0.074*** (t = 7.212)	0.077*** (t = 7.476)
HOT:pm1730	0.209*** (t = 20.459)	0.211*** (t = 20.594)	0.205*** (t = 20.018)	0.208*** (t = 20.294)
HOT:pm1800	0.406*** (t = 40.890)	0.409*** (t = 41.100)	0.402*** (t = 40.470)	0.405*** (t = 40.754)
HOT:pm1830	0.346*** (t = 35.484)	0.348*** (t = 35.718)	0.343*** (t = 35.247)	0.345*** (t = 35.447)
HOT Share	0.5407	0.5407	0.5407	0.5407
Observations	1,191,877	1,191,877	1,191,877	1,191,877
R^2	0.215	0.215	0.214	0.215
Log Likelihood	-645,007.40	-644,911.30	-645,784.60	-645,304.40

*p<0.1; **p<0.05; ***p<0.01

Model 15 – Toll over Segment Count

The next pair of models includes the toll amount divided by the number of segments traversed, to investigate whether users consider toll rate as a function of trip length in their lane choice decision making. Table 93 presents the results from the AM peak period data set. While the tollSegments interaction term is significant at the 95% confidence level, and yields the highest test statistic in the model, the overall goodness of fit suffers relative to the Model 14 variant. The odds ratio for the morning peak tollSegments term is 0.0930, indicating a large decrease in toll lane use probability with a unit increase in the toll amount divided by the segment count.

Table 93: Toll over Segment Count - AM Peak Model

	AM Peak – Model 15
Intercept	-1.232*** (t = -30.814)
avgSpeed^2	-0.0003*** (t = -105.898)
tollSegments	-2.375*** (t = -321.387)
transponderCount	-0.002*** (t = -87.872)
HOT: congested50	1.816*** (t = 242.420)
HOT: hhEdu	-0.173*** (t = -58.159)
HOT: hhAge	-0.024*** (t = -13.007)
HOT: log(hhIncomeDollars)	0.195*** (t = 50.244)
HOT: hhSize	-0.034*** (t = -27.160)
HOT: february	0.263*** (t = 25.584)
HOT: march	0.218*** (t = 21.897)
HOT: april	0.365*** (t = 36.022)
HOT: may	0.364*** (t = 36.463)
HOT: june	0.239*** (t = 22.966)
HOT: july	0.132*** (t = 13.172)
HOT: august	0.385*** (t = 38.022)
HOT: september	0.538*** (t = 51.874)
HOT: october	0.538*** (t = 52.097)
HOT: november	0.384*** (t = 36.761)
HOT: december	0.163*** (t = 15.673)
HOT: tuesday	0.246*** (t = 37.180)
HOT: wednesday	0.246*** (t = 36.745)
HOT: thursday	0.240*** (t = 36.064)
HOT: friday	-0.827*** (t = -112.363)
HOT:am630	1.279*** (t = 141.239)
HOT: am700	1.523*** (t = 156.005)
HOT: am730	1.628*** (t = 161.653)
HOT:am800	1.513*** (t = 151.972)
HOT:am830	1.265*** (t = 129.947)
HOT:am900	0.786*** (t = 82.651)
HOT:am930	0.024** (t = 2.481)
HOT Share	0.5278
Observations	1,105,171
R^2	0.128
Log Likelihood	-666,537.80

*p<0.1; **p<0.05; ***p<0.01

Table 94 presents the results from the PM peak trips with the tollSegments interaction term included. Here the coefficient is again negative and significant at the 95% confidence level, though the t-statistic is not the highest one present. Similar to the AM peak model in the previous table, the model goodness of fit measure suffers when the toll amount and segment count factors are replaced by the single interaction term. The odds ratio for the tollSegments

term in the afternoon peak is closer to a value of one than in the morning peak (0.355 for the PM peak, 0.0930 for the AM peak).

Table 94: Toll over Segment Count - PM Peak Model

	PM Peak – Model 15
Intercept	-1.817*** (t = -47.329)
avgSpeed^2	0.0002*** (t = 66.464)
tollSegments	-1.036*** (t = -118.903)
transponderCount	0.001*** (t = 35.538)
HOT: congested40	1.316*** (t = 260.457)
HOT: hhEdu	-0.255*** (t = -90.585)
HOT: hhAge	-0.002 (t = -0.963)
HOT: log(hhIncomeDollars)	0.200*** (t = 54.624)
HOT: hhSize	-0.009*** (t = -7.237)
HOT: february	-0.033*** (t = -3.430)
HOT: march	0.005 (t = 0.501)
HOT: april	-0.011 (t = -1.143)
HOT: may	-0.074*** (t = -7.910)
HOT: june	-0.003 (t = -0.293)
HOT: july	-0.024** (t = -2.462)
HOT: august	0.014 (t = 1.455)
HOT: september	0.157*** (t = 15.937)
HOT: october	0.180*** (t = 18.463)
HOT: november	0.110*** (t = 11.142)
HOT: december	0.056*** (t = 5.671)
HOT: tuesday	0.043*** (t = 7.072)
HOT: wednesday	0.066*** (t = 10.578)
HOT: thursday	0.040*** (t = 6.193)
HOT: friday	0.070*** (t = 10.240)
HOT:pm1530	0.172*** (t = 19.207)
HOT:pm1600	0.233*** (t = 26.017)
HOT:pm1630	0.321*** (t = 35.331)
HOT:pm1700	0.344*** (t = 37.468)
HOT:pm1730	0.339*** (t = 36.949)
HOT:pm1800	0.388*** (t = 43.170)
HOT:pm1830	0.303*** (t = 34.130)
HOT Share	0.5407
Observations	1,191,877
R^2	0.09
Log Likelihood	-748,055.50

*p<0.1; **p<0.05; ***p<0.01

Model 16 – All Interaction Terms

The next set of models includes all of the interaction terms from this section: toll amount over segment count, toll over income, and income over household size. Model 16a in the AM and PM peak excludes the household income and size measures outside of the interaction terms, while Model 16b includes both measures alongside their respective interaction terms. The inclusion of those variables in Model 16b changes the sign of the intercept term, and yields a significant coefficient for the household age factor (though the magnitude of the coefficient is still small). The pseudo-R^2 goodness of fit measure improves by 0.002.

The toll over log(income) coefficient is positive now, while previously in Model 12 the estimator was negative. The income over household size coefficient remains positive and significant as it was in previous models; the tollSegments interaction term also remains negative and significant at the 95% confidence level. The change in the tollLogIncome term may be due to correlation with other interaction terms, as the models now have multiple instances of both toll and household income represented in their utility equations. In both 16a and 16b, the goodness of fit measures are lower than their previous peaks in Model 14.

Table 95: All Interaction Terms - AM Peak Models

	AM Peak – Model 16a	AM Peak – Model 16b
Intercept	0.322*** (t = 16.653)	-1.926*** (t = -43.651)
$avgSpeed^2$	-0.0003*** (t = -98.561)	-0.0003*** (t = -99.147)
tollSegments	-2.815*** (t = -272.695)	-2.871*** (t = -275.551)
tollLogIncome	1.527*** (t = 64.643)	1.692*** (t = 70.898)
transponderCount	-0.001*** (t = -69.873)	-0.001*** (t = -68.576)
HOT: congested50	1.695*** (t = 221.016)	1.685*** (t = 219.323)
HOT:I(log(hhIncomeDollars)/hhSize)	0.022*** (t = 14.852)	0.055*** (t = 15.583)
HOT: log(hhIncomeDollars)		0.221*** (t = 56.173)
HOT: hhSize		0.004 (t = 1.547)
HOT: hhEdu	-0.106*** (t = -37.580)	-0.162*** (t = -54.274)
HOT: hhAge	-0.001 (t = -0.719)	-0.022*** (t = -11.773)
HOT: february	0.252*** (t = 24.511)	0.253*** (t = 24.587)
HOT: march	0.210*** (t = 21.003)	0.212*** (t = 21.187)
HOT: april	0.352*** (t = 34.751)	0.355*** (t = 34.999)
HOT: may	0.359*** (t = 35.910)	0.362*** (t = 36.125)
HOT: june	0.236*** (t = 22.677)	0.237*** (t = 22.745)
HOT: july	0.127*** (t = 12.726)	0.126*** (t = 12.603)
HOT: august	0.385*** (t = 38.039)	0.385*** (t = 37.931)
HOT: september	0.547*** (t = 52.608)	0.549*** (t = 52.698)
HOT: october	0.549*** (t = 53.087)	0.552*** (t = 53.260)
HOT: november	0.398*** (t = 38.005)	0.400*** (t = 38.186)
HOT: december	0.167*** (t = 16.046)	0.168*** (t = 16.090)
HOT: tuesday	0.242*** (t = 36.486)	0.244*** (t = 36.749)
HOT: wednesday	0.238*** (t = 35.597)	0.240*** (t = 35.821)
HOT: thursday	0.233*** (t = 34.968)	0.235*** (t = 35.246)
HOT: friday	-0.813*** (t = -110.564)	-0.812*** (t = -110.273)
HOT:am630	1.245*** (t = 137.649)	1.247*** (t = 137.644)
HOT: am700	1.536*** (t = 157.368)	1.545*** (t = 157.901)
HOT: am730	1.634*** (t = 162.341)	1.643*** (t = 162.927)
HOT:am800	1.501*** (t = 150.934)	1.511*** (t = 151.688)
HOT:am830	1.267*** (t = 130.315)	1.277*** (t = 131.106)
HOT:am900	0.795*** (t = 83.878)	0.801*** (t = 84.442)
HOT:am930	0.052*** (t = 5.295)	0.057*** (t = 5.841)
HOT Share	0.5278	0.5278
Observations	1,105,171	1,105,171
R^2	0.129	0.131
Log Likelihood	-665,561.20	-663,836.90

*p<0.1; **p<0.05; ***p<0.01

The PM peak models with all interaction terms included are presented in Table 96. As in the AM peak results, the goodness of fit measures suffer relative to Model 14. The tollLogIncome term, which divides the toll about by the log of the household's income, is once again positive where previously its coefficient was negative. The coefficient of household income divided by household size exhibits more complex behavior; where it was previously

positive, it is now negative in the absence of separate household income and size terms and positive when those terms are included. As in Model 15, the tollSegments term remains negative and significant at the 95% confidence level.

Table 96: All Interaction Terms - PM Peak Models

	PM Peak – Model 16a	PM Peak – Model 16b
Intercept	0.775*** (t = 38.482)	-2.425*** (t = -55.509)
$avgSpeed^2$	-0.00002*** (t = -4.250)	-0.00002*** (t = -5.838)
tollSegments	-5.164*** (t = -275.038)	-5.336*** (t = -279.724)
tollLogIncome	10.822*** (t = 263.025)	11.211*** (t = 269.019)
transponderCount	0.003*** (t = 106.525)	0.003*** (t = 109.778)
HOT: congested40	1.309*** (t = 252.010)	1.325*** (t = 253.890)
HOT:I(log(hhIncomeDollars)/hhSize)	-0.017*** (t = -11.723)	0.015*** (t = 4.329)
HOT: log(hhIncomeDollars)		0.328*** (t = 84.374)
HOT: hhSize		-0.007*** (t = -2.650)
HOT: hhEdu	-0.133*** (t = -48.080)	-0.215*** (t = -73.396)
HOT: hhAge	0.028*** (t = 15.862)	-0.002 (t = -0.871)
HOT: february	-0.056*** (t = -5.659)	-0.058*** (t = -5.808)
HOT: march	-0.029*** (t = -2.951)	-0.031*** (t = -3.099)
HOT: april	-0.109*** (t = -11.025)	-0.111*** (t = -11.149)
HOT: may	-0.189*** (t = -19.481)	-0.194*** (t = -19.993)
HOT: june	-0.138*** (t = -13.806)	-0.145*** (t = -14.474)
HOT: july	-0.147*** (t = -14.927)	-0.156*** (t = -15.737)
HOT: august	-0.071*** (t = -7.169)	-0.077*** (t = -7.816)
HOT: september	0.182*** (t = 17.810)	0.183*** (t = 17.762)
HOT: october	0.202*** (t = 19.920)	0.204*** (t = 20.001)
HOT: november	0.133*** (t = 12.991)	0.135*** (t = 13.115)
HOT: december	0.107*** (t = 10.466)	0.111*** (t = 10.854)
HOT: tuesday	0.022*** (t = 3.555)	0.021*** (t = 3.360)
HOT: wednesday	0.031*** (t = 4.841)	0.028*** (t = 4.398)
HOT: thursday	-0.075*** (t = -11.097)	-0.084*** (t = -12.392)
HOT: friday	-0.055*** (t = -7.666)	-0.064*** (t = -8.931)
HOT:pm1530	0.016* (t = 1.724)	0.013 (t = 1.445)
HOT:pm1600	-0.030*** (t = -3.289)	-0.038*** (t = -4.075)
HOT:pm1630	0.030*** (t = 3.131)	0.024** (t = 2.499)
HOT:pm1700	0.069*** (t = 7.254)	0.065*** (t = 6.787)
HOT:pm1730	0.137*** (t = 14.375)	0.137*** (t = 14.306)
HOT:pm1800	0.290*** (t = 31.317)	0.295*** (t = 31.830)
HOT:pm1830	0.270*** (t = 29.823)	0.277*** (t = 30.454)
HOT Share	0.5407	0.5407
Observations	1,191,877	1,191,877
R^2	0.139	0.144
Log Likelihood	-707,686.40	-703,900.70

*p<0.1; **p<0.05; ***p<0.01

Model 17 – Additional Interaction Term Combinations

The final set of standard binary logit models presented here investigates different combinations of the interaction terms, rather than looking at each in isolation and then all together. Of the three combinations represented in Table 97, the best-performing model in the AM peak is Model 17a, in which the toll over income and income over household interaction terms are both included. The coefficients of all three models are uniformly significant at the 95% confidence level; beyond the constants, there are no differences in coefficient signs and only small differences in magnitudes.

Table 97: Additional Interaction Term Combinations - AM Peak

	AM Peak – Model 17a	AM Peak – Model 17b	AM Peak – Model 17c
Intercept	1.402*** (t = 19.709)	-1.456*** (t = -20.910)	0.852*** (t = 12.150)
$avgSpeed^2$	-0.0004*** (t = -114.857)	-0.0004*** (t = -110.164)	-0.0003*** (t = -99.913)
tollLogIncome	-7.438*** (t = -330.151)		
$tollAmount^2$		-0.067*** (t = -287.831)	
tollSegments			-2.315*** (t = -304.822)
transponderCount	-0.001*** (t = -41.305)	-0.001*** (t = -58.875)	-0.001*** (t = -39.451)
HOT: congested50	1.596*** (t = 202.689)	1.345*** (t = 178.022)	1.514*** (t = 196.218)
HOT: hhEdu	-0.136*** (t = -44.596)	-0.136*** (t = -45.249)	-0.141*** (t = -46.396)
HOT: hhAge	-0.031*** (t = -15.921)	-0.030*** (t = -15.652)	-0.030*** (t = -15.478)
HOT:I(hhIncomeDollars/hhSize)	0.00002*** (t = 58.585)	0.00002*** (t = 52.678)	0.00002*** (t = 53.735)
HOT: log(hhIncomeDollars)	-0.448*** (t = -58.573)	-0.202*** (t = -27.077)	-0.193*** (t = -25.770)
HOT: hhSize	0.061*** (t = 28.070)	0.046*** (t = 21.870)	0.054*** (t = 25.161)
HOT: segmentCount	0.925*** (t = 338.884)	0.759*** (t = 306.294)	0.348*** (t = 166.440)
HOT: February	0.264*** (t = 24.968)	0.225*** (t = 21.742)	0.247*** (t = 23.720)
HOT: march	0.234*** (t = 22.859)	0.188*** (t = 18.722)	0.216*** (t = 21.294)
HOT: april	0.378*** (t = 36.341)	0.293*** (t = 28.766)	0.355*** (t = 34.538)
HOT: may	0.376*** (t = 36.683)	0.312*** (t = 31.064)	0.366*** (t = 36.095)
HOT: june	0.272*** (t = 25.447)	0.238*** (t = 22.703)	0.251*** (t = 23.751)
HOT: july	0.147*** (t = 14.237)	0.121*** (t = 11.950)	0.130*** (t = 12.801)
HOT: august	0.409*** (t = 39.418)	0.339*** (t = 33.280)	0.397*** (t = 38.625)
HOT: September	0.580*** (t = 54.547)	0.524*** (t = 50.220)	0.576*** (t = 54.602)
HOT: October	0.578*** (t = 54.715)	0.520*** (t = 50.039)	0.582*** (t = 55.428)
HOT: November	0.447*** (t = 41.644)	0.403*** (t = 38.230)	0.440*** (t = 41.409)
HOT: December	0.213*** (t = 19.842)	0.172*** (t = 16.345)	0.193*** (t = 18.252)
HOT: Tuesday	0.237*** (t = 34.968)	0.242*** (t = 36.350)	0.240*** (t = 35.648)
HOT: Wednesday	0.242*** (t = 35.418)	0.249*** (t = 36.993)	0.236*** (t = 34.785)
HOT: Thursday	0.235*** (t = 34.550)	0.242*** (t = 36.170)	0.231*** (t = 34.190)
HOT: Friday	-0.875*** (t = -115.484)	-0.620*** (t = -85.434)	-0.821*** (t = -110.165)
HOT:am630	1.579*** (t = 164.734)	1.248*** (t = 136.562)	1.339*** (t = 145.856)
HOT: am700	1.920*** (t = 186.087)	1.588*** (t = 160.497)	1.717*** (t = 171.632)
HOT: am730	1.975*** (t = 187.125)	1.636*** (t = 161.781)	1.791*** (t = 174.333)
HOT:am800	1.830*** (t = 175.190)	1.515*** (t = 151.013)	1.632*** (t = 160.980)
HOT:am830	1.583*** (t = 154.540)	1.352*** (t = 136.143)	1.404*** (t = 141.334)
HOT:am900	1.113*** (t = 111.193)	1.116*** (t = 114.249)	0.929*** (t = 96.106)
HOT:am930	0.364*** (t = 35.193)	0.788*** (t = 79.057)	0.204*** (t = 20.441)
HOT Share	0.5278	0.5278	0.5278
Observations	1,105,171	1,105,171	1,105,171
R^2	0.163	0.138	0.149
Log Likelihood	-639,913.70	-658,649.00	-650,476.70

*p<0.1; **p<0.05; ***p<0.01

Table 98 presents similar results for the interaction term combinations in the PM peak. Again, Model 17a outperforms the other two in the R^2 and log-likelihood measures. The consistency of the estimated coefficients is not as strong as in the morning period models: the

day of week and time of day estimators in the PM peak see more variation in their magnitudes and significance levels.

Table 98: Additional Interaction Term Combinations - PM Peak

	PM Peak – Model 17a	PM Peak – Model 17b	PM Peak – Model 17c
Intercept	-0.941*** (t = -13.372)	-1.723*** (t = -24.538)	-1.095*** (t = -15.619)
$avgSpeed^2$	-0.0002*** (t = -46.743)	-0.0002*** (t = -43.481)	-0.0002*** (t = -47.882)
tollLogIncome	-5.036*** (t = -193.992)	()	()
$tollAmount^2$	()	-0.047*** (t = -151.644)	()
tollSegments	()	()	-1.589*** (t = -160.935)
transponderCount	0.004*** (t = 139.749)	0.004*** (t = 156.244)	0.004*** (t = 150.407)
HOT: congested40	1.363*** (t = 246.814)	1.231*** (t = 229.927)	1.326*** (t = 240.372)
HOT: hhEdu	-0.212*** (t = -68.456)	-0.209*** (t = -67.876)	-0.204*** (t = -66.154)
HOT: hhAge	-0.007*** (t = -3.559)	-0.006*** (t = -3.403)	-0.006*** (t = -3.288)
HOT:I(hhIncomeDollars/hhSize)	0.00001*** (t = 30.051)	0.00001*** (t = 27.708)	0.00001*** (t = 27.845)
HOT: log(hhIncomeDollars)	-0.152*** (t = -20.224)	-0.050*** (t = -6.715)	-0.055*** (t = -7.345)
HOT: hhSize	0.032*** (t = 14.700)	0.026*** (t = 12.357)	0.026*** (t = 12.175)
HOT: segmentCount	1.096*** (t = 406.884)	0.953*** (t = 400.764)	0.866*** (t = 391.825)
HOT: february	-0.072*** (t = -6.695)	-0.087*** (t = -8.164)	-0.077*** (t = -7.298)
HOT: march	-0.059*** (t = -5.564)	-0.109*** (t = -10.333)	-0.070*** (t = -6.671)
HOT: april	-0.275*** (t = -25.888)	-0.306*** (t = -29.021)	-0.263*** (t = -24.996)
HOT: may	-0.375*** (t = -35.867)	-0.391*** (t = -37.663)	-0.361*** (t = -34.869)
HOT: june	-0.372*** (t = -34.469)	-0.374*** (t = -34.940)	-0.350*** (t = -32.762)
HOT: july	-0.359*** (t = -33.724)	-0.371*** (t = -35.080)	-0.343*** (t = -32.559)
HOT: august	-0.156*** (t = -14.801)	-0.269*** (t = -25.882)	-0.180*** (t = -17.245)
HOT: september	0.239*** (t = 22.130)	0.048*** (t = 4.543)	0.189*** (t = 17.476)
HOT: october	0.279*** (t = 26.046)	0.075*** (t = 7.175)	0.223*** (t = 20.814)
HOT: november	0.221*** (t = 20.393)	0.047*** (t = 4.421)	0.170*** (t = 15.727)
HOT: december	0.203*** (t = 18.678)	0.035*** (t = 3.290)	0.158*** (t = 14.650)
HOT: tuesday	0.016** (t = 2.442)	0.001 (t = 0.146)	0.009 (t = 1.399)
HOT: wednesday	0.027*** (t = 3.986)	-0.004 (t = -0.597)	0.012* (t = 1.779)
HOT: thursday	-0.110*** (t = -15.404)	-0.182*** (t = -25.713)	-0.146*** (t = -20.509)
HOT: friday	-0.089*** (t = -11.685)	-0.170*** (t = -22.498)	-0.130*** (t = -17.113)
HOT:pm1530	-0.057*** (t = -5.690)	-0.141*** (t = -14.353)	-0.089*** (t = -9.059)
HOT:pm1600	-0.139*** (t = -13.874)	-0.303*** (t = -30.737)	-0.201*** (t = -20.264)
HOT:pm1630	-0.018* (t = -1.728)	-0.264*** (t = -26.398)	-0.113*** (t = -11.167)
HOT:pm1700	0.072*** (t = 7.008)	-0.210*** (t = -21.092)	-0.038*** (t = -3.751)
HOT:pm1730	0.202*** (t = 19.765)	-0.097*** (t = -9.837)	0.089*** (t = 8.820)
HOT:pm1800	0.400*** (t = 40.324)	0.126*** (t = 13.214)	0.306*** (t = 31.078)
HOT:pm1830	0.342*** (t = 35.149)	0.150*** (t = 15.824)	0.288*** (t = 29.936)
HOT Share	0.5407	0.5407	0.5407
Observations	1,191,877	1,191,877	1,191,877
R^2	0.215	0.205	0.208
Log Likelihood	-645,127.80	-653,163.00	-651,317.80

*p<0.1; **p<0.05; ***p<0.01

Overview of Interaction Term Investigation

Table 99 presents an overview of the model results for the interaction term investigation. The cells with highlighted text represent the models with the best goodness-of-fit or AIC measures within that category. Unlike the previous model comparison, here the best performing models in the morning and afternoon peaks are the same. Model 14b has the best goodness of fit measure and the lowest AIC value in both time frames. In both cases, Model 17a is a close second, however the additional interaction term in 17a does not provide enough of a benefit to warrant its inclusion. This effect is even more apparent in Models 16a and 16b, in which all of the interaction terms were included to little benefit. For this reason, Model 14b will be the model of choice going forwards.

Table 99: Summary of Models with Additional Interaction Terms

Model	AM Peak - R^2 Value	AM Peak – AIC	PM Peak – R^2 Value	PM Peak – AIC
Model 12a	0.161	1,366,711	0.214	1,294,997
Model 12b	0.161	1,366,405	0.214	1,294,907
Model 13a	0.081	1,497,011	0.193	1,329,590
Model 13b	0.082	1,494,839	0.194	1,329,255
Model 14a	0.164	1,361,665	0.215	1,293,750
Model 14b	0.164	1,360,791	0.215	1,293,560
Model 14c	0.162	1,364,956	0.214	1,295,305
Model 14d	0.163	1,363,758	0.215	1,294,349
Model 15	0.125	1,424,731	0.090	1,500,244
Model 16a	0.128	1,420,074	0.139	1,418,996
Model 16b	0.130	1,416,159	0.144	1,411,425
Model 17a	0.163	1,362,616	0.215	1,293,991
Model 17b	0.138	1,403,695	0.205	1,310,056
Model 17c	0.150	1,384,399	0.208	1,306,313

Income Segmentation

As in the previously published paper on the initial modeling work, a primary goal of this dissertation is to investigate differences in lane use decision making among different income segments. For that reason, the previously selected Model 14b is estimated below for the Lower,

Medium, and Higher income groups. Like before, those categories are based on the annual household income measure in the Epsilon demographic data: 'Lower' income households make less than $50,000 annually, 'Medium' income households make between $50,000 and $100,000, and 'Higher' income households make over $100,000. The overview of the different income segments can be found earlier in this chapter in Table 70.

Three Income Segments

Table 100 presents the results from estimating Model 14b for the AM peak period in calendar year 2013. The first difference can be seen in the intercept terms: only the medium income segment has a positive coefficient, and the higher income coefficient is much larger in magnitude than either of the others. This is perhaps counterintuitive as the higher income segment has the largest share of HOT trips. The coefficients for the transponder count difference also vary in their signs and magnitudes: the lower income segment has the largest absolute coefficient, while the higher income segment has the only positive estimator. In all cases the values are very close to zero, however.

The household age coefficients are all negative, but their estimators differ by orders of magnitude. The medium income segment age coefficient is the smallest (and fails to achieve significance at the 95% confidence level), followed by the lower income segment coefficient. The higher income segment has the largest coefficient with the greatest test statistic. To the extent that income increases with age, this may reflect retirement-aged households. As mentioned previously, household age, size, and education are all positively correlated with household income.

Household size also has different impacts on the three different income segments: larger households increase toll lane choice probability in the middle income group, but decrease that

probability for the lower and higher segments. The interaction term included here, which divides household income by size, is very close to zero. It fails to achieve 95% significance for the lower income segment, where the estimate is zero.

The household income coefficients are all significant at the 95% confidence level, but vary in their signs and magnitude. The lower and medium segment estimators are both negative, indicating that within their income spectrums, a higher household income reduces the probability of selecting the toll lane on a given trip. The higher income segment, however, yields an income estimator that is positive and larger than its counterparts. This effect was also present in the initial modeling analysis: the higher income segment had a positive household income coefficient which was larger in magnitude than those of the lower and medium segments. As discussed in that chapter as well, that may be due to the greater diversity of household incomes within that category. The lower income segment includes five distinct income values, while the medium segment includes two. In contrast, the higher income segment contains eight. This difference in the income effects among the higher income segment inspired further investigation which is described in the next part of this chapter.

The estimated coefficients for the month and time of day factors are consistent across all three income segments. Coefficient signs are the same, and the magnitudes are likewise very similar. In all three segments, trips in the 7:30 – 7:59 AM interval see the highest increase in Express Lane choice probability. Likewise, Friday trips are the only trips for all three groups that see a decrease in toll lane probability relative to Monday trips. This effect is also evident in the monthly coefficients: October yields the largest estimator for all three groups.

Also of interest are the similarities between the three models. Toll amount, for example, yields coefficients that are close in sign and magnitude across all three income categories.

Similarly, the (square of) speed difference coefficients are all negative and very small, while the congestion dummy coefficients are all positive with magnitudes that differ by 0.036. Household education is uniformly negative across all segments; again, this factor is positively correlated with income. Finally, the segmentCount variable, which was designed to replace distance, is positive and highly significant in all three models. Goodness of fit, as measured by the McFadden's pseudo-R^2 value, differ by only 0.05. Note that McFadden's pseudo-R^2 measures the log likelihood of the full model against the log likelihood of the intercept-only model.

Table 100: Model 14b with 3 Income Segments - AM Peak

	AM Peak – Lower	AM Peak – Medium	AM Peak – Higher
Intercept	-1.247*** (t = -8.971)	2.328*** (t = 9.177)	-11.780*** (t = -35.412)
$avgSpeed^2$	-0.0003*** (t = -51.375)	-0.0003*** (t = -80.858)	-0.0004*** (t = -64.548)
tollAmount	-0.694*** (t = -183.049)	-0.701*** (t = -237.269)	-0.668*** (t = -169.961)
transponderCount	-0.001*** (t = -37.234)	-0.0002*** (t = -5.800)	0.0001*** (t = 2.927)
HOT: congested50	1.446*** (t = 105.719)	1.482*** (t = 133.393)	1.483*** (t = 99.224)
HOT: hhEdu	-0.142*** (t = -27.601)	-0.110*** (t = -24.393)	-0.126*** (t = -18.818)
HOT: hhAge	-0.030*** (t = -9.138)	-0.001 (t = -0.479)	-0.114*** (t = -25.670)
HOT:I(hhIncomeDollars)/hhSize)	0 (t = -0.725)	0.00002*** (t = 26.440)	0.00001*** (t = 11.326)
HOT: log(hhIncomeDollars)	-0.177*** (t = -11.353)	-0.545*** (t = -22.225)	0.776*** (t = 25.216)
HOT: hhSize	-0.056*** (t = -10.499)	0.045*** (t = 11.012)	-0.010** (t = -2.006)
HOT: segmentCount	0.989*** (t = 199.890)	0.960*** (t = 242.991)	0.952*** (t = 178.977)
HOT: february	0.310*** (t = 16.187)	0.245*** (t = 16.182)	0.242*** (t = 12.075)
HOT: march	0.240*** (t = 12.896)	0.253*** (t = 17.138)	0.243*** (t = 12.411)
HOT: april	0.413*** (t = 21.942)	0.398*** (t = 26.508)	0.384*** (t = 19.261)
HOT: may	0.387*** (t = 20.930)	0.379*** (t = 25.710)	0.405*** (t = 20.704)
HOT: june	0.255*** (t = 13.170)	0.319*** (t = 20.738)	0.281*** (t = 13.728)
HOT: july	0.101*** (t = 5.416)	0.198*** (t = 13.368)	0.139*** (t = 7.055)
HOT: august	0.403*** (t = 21.405)	0.444*** (t = 29.853)	0.419*** (t = 21.174)
HOT: september	0.586*** (t = 30.504)	0.633*** (t = 41.392)	0.625*** (t = 30.807)
HOT: october	0.587*** (t = 30.685)	0.653*** (t = 42.934)	0.639*** (t = 31.635)
HOT: november	0.408*** (t = 20.960)	0.549*** (t = 35.634)	0.518*** (t = 25.109)
HOT: december	0.172*** (t = 8.884)	0.262*** (t = 17.094)	0.305*** (t = 14.877)
HOT: tuesday	0.290*** (t = 23.706)	0.209*** (t = 21.388)	0.225*** (t = 17.418)
HOT: Wednesday	0.289*** (t = 23.420)	0.203*** (t = 20.588)	0.228*** (t = 17.392)
HOT: Thursday	0.276*** (t = 22.400)	0.211*** (t = 21.490)	0.197*** (t = 15.065)
HOT: Friday	-0.924*** (t = -67.837)	-0.917*** (t = -84.405)	-0.897*** (t = -61.378)
HOT:am630	1.686*** (t = 98.432)	1.579*** (t = 117.295)	1.551*** (t = 83.740)
HOT: am700	2.061*** (t = 111.143)	1.898*** (t = 129.459)	1.975*** (t = 99.261)
HOT: am730	2.190*** (t = 115.371)	1.906*** (t = 127.507)	2.041*** (t = 100.646)
HOT:am800	2.071*** (t = 110.020)	1.805*** (t = 122.752)	1.787*** (t = 88.583)
HOT:am830	1.843*** (t = 99.346)	1.523*** (t = 104.676)	1.501*** (t = 76.157)
HOT:am900	1.237*** (t = 67.561)	1.033*** (t = 73.026)	1.060*** (t = 54.549)
HOT:am930	0.375*** (t = 19.969)	0.283*** (t = 19.350)	0.257*** (t = 12.775)
HOT Share	0.5068	0.5202	0.5562
Observations	342,209	533,623	301,182
R^2	0.171	0.166	0.166
Log Likelihood	-196,702.70	-308,201.30	-172,527.40

*p<0.1; **p<0.05; ***p<0.01

Table 101 includes the results from the PM peak model estimates across the three income segments. Once again, the first difference is apparent in the intercept term, for which the higher income segment coefficient has the largest, most negative magnitude. In fact the discrepancies are similar to those found in the morning peak models: the household age, income, and size factors have the most notable differences. As in the AM peak models, household age yields a

negative coefficient only in the highest income segment. That segment is also unique in having a positive coefficient for household income, just like the morning trips.

The model similarities also mirror those of the morning period models, starting with the effects of the speed difference between the lanes. As was the case previously, each segment coefficient was negative, very small, and significant at the 99% confidence level. The transponderCount coefficients are once again very close to zero, though their test statistics indicate high levels of significance. Toll level coefficients are negative and in the 0.4 to 0.5 range for all three models. The congested conditions dummy coefficients are all positive, significant, and similarly-sized, while the household education coefficients are all negative, significant, and similarly-sized.

Afternoon toll lane trips exhibit a pattern in the monthly, daily, and half-hour dummy coefficients. Corridor users in the afternoon are less likely to take toll lane trips between February and August, relative to their January probability. From September through December, positive coefficients indicate higher HOT probabilities relative to January. Similarly, Tuesday and Wednesday trips are, all else being equal, more likely to be toll lane trips compared to those taken on Monday. That relationship is inverted on Thursday and Friday for all segments. This effect is also apparent in the time of day factors: positive coefficients appear only after 5pm (4:30pm for the higher income segment).

The model goodness of fit measures, represented here as elsewhere by McFadden's pseudo-R^2, exhibit an inverse pattern relative to the morning trips. Here the lower income segment has the model with the lowest pseudo-R^2 value, while the higher income segment has the highest. All three models outperform the morning peak models, a result that has been consistent throughout the dissertation.

Table 101: Model 14b with 3 Income Segments - PM Peak

	PM Peak – Lower	PM Peak – Medium	PM Peak – Higher
Intercept	-1.142*** (t = -8.035)	-1.128*** (t = -4.261)	-4.002*** (t = -11.638)
avgSpeed^2	-0.0002*** (t = -34.684)	-0.0002*** (t = -29.407)	-0.0001*** (t = -14.248)
tollAmount	-0.431*** (t = -98.171)	-0.467*** (t = -136.133)	-0.476*** (t = -99.870)
transponderCount	0.003*** (t = 61.096)	0.004*** (t = 99.784)	0.004*** (t = 79.693)
HOT: congested40	1.367*** (t = 135.420)	1.344*** (t = 165.121)	1.410*** (t = 124.823)
HOT: hhEdu	-0.200*** (t = -38.167)	-0.192*** (t = -40.699)	-0.299*** (t = -42.010)
HOT: hhAge	0.012*** (t = 3.759)	0.001 (t = 0.512)	-0.077*** (t = -16.555)
HOT:I(hhIncomeDollars)/hhSize)	0.00000** (t = 1.985)	0.00000*** (t = 4.717)	0.00001*** (t = 13.437)
HOT: log(hhIncomeDollars)	-0.111*** (t = -7.005)	-0.112*** (t = -4.359)	0.148*** (t = 4.647)
HOT: hhSize	-0.013** (t = -2.462)	-0.020*** (t = -4.617)	0.047*** (t = 9.139)
HOT: segmentCount	1.068*** (t = 218.616)	1.079*** (t = 272.467)	1.157*** (t = 207.768)
HOT: february	-0.064*** (t = -3.265)	-0.067*** (t = -4.286)	-0.095*** (t = -4.369)
HOT: march	-0.042** (t = -2.160)	-0.054*** (t = -3.484)	-0.090*** (t = -4.171)
HOT: april	-0.250*** (t = -12.897)	-0.272*** (t = -17.353)	-0.323*** (t = -14.851)
HOT: may	-0.376*** (t = -19.675)	-0.356*** (t = -23.082)	-0.408*** (t = -19.130)
HOT: june	-0.395*** (t = -20.012)	-0.333*** (t = -20.980)	-0.402*** (t = -18.182)
HOT: july	-0.366*** (t = -18.800)	-0.314*** (t = -19.978)	-0.438*** (t = -20.109)
HOT: august	-0.184*** (t = -9.500)	-0.097*** (t = -6.282)	-0.215*** (t = -10.054)
HOT: september	0.139*** (t = 7.012)	0.328*** (t = 20.554)	0.221*** (t = 10.049)
HOT: october	0.178*** (t = 9.098)	0.375*** (t = 23.748)	0.250*** (t = 11.440)
HOT: november	0.092*** (t = 4.621)	0.325*** (t = 20.300)	0.219*** (t = 9.929)
HOT: december	0.118*** (t = 5.971)	0.266*** (t = 16.593)	0.212*** (t = 9.576)
HOT: tuesday	0.024* (t = 1.954)	0.012 (t = 1.198)	0.011 (t = 0.775)
HOT: wednesday	0.049*** (t = 3.960)	0.022** (t = 2.221)	0.003 (t = 0.246)
HOT: thursday	-0.069*** (t = -5.316)	-0.117*** (t = -11.095)	-0.149*** (t = -10.211)
HOT: friday	-0.031** (t = -2.261)	-0.120*** (t = -10.640)	-0.106*** (t = -6.772)
HOT:pm1530	-0.001 (t = -0.037)	-0.058*** (t = -3.861)	-0.122*** (t = -6.199)
HOT:pm1600	-0.108*** (t = -5.874)	-0.156*** (t = -10.379)	-0.128*** (t = -6.395)
HOT:pm1630	-0.068*** (t = -3.597)	-0.037** (t = -2.429)	0.115*** (t = 5.615)
HOT:pm1700	0.019 (t = 0.984)	0.088*** (t = 5.728)	0.158*** (t = 7.691)
HOT:pm1730	0.099*** (t = 5.313)	0.229*** (t = 15.018)	0.329*** (t = 16.191)
HOT:pm1800	0.288*** (t = 15.807)	0.418*** (t = 28.278)	0.548*** (t = 27.612)
HOT:pm1830	0.241*** (t = 13.371)	0.344*** (t = 23.759)	0.501*** (t = 25.853)
HOT Share	0.5223	0.5389	0.5684
Observations	349,783	544,660	300,556
R^2	0.199	0.213	0.242
Log Likelihood	-193,922.40	-295,646.70	-155,718.00

*p<0.1; **p<0.05; ***p<0.01

Five Income Segments

The investigation of the three-segment lane choice modeling strategy in the previous section showed that in two categories, namely the constant term and the household income coefficient, the higher income segment differed from the other two in its results. This prompted the question of whether the segmentation strategy was minimizing behavioral differences among the higher income segment by grouping them together. For that reason, this section examines the model

selected in this chapter (Model 14b) with five different income segments rather than three. The purpose is to investigate the behavior of the users at the highest end of the income spectrum to see if more variability is present in their decisions.

Table 102 provides an overview of the five income segments. Segments A and B are identical to the Lower and Medium household income segments of the previous sections. Segments C through E further subdivide the Higher income segment. Segment C represents annual household incomes of $100-149k, Segment D includes households with annual incomes from $150-199k, and Segment E is populated by those households with $200k and more in annual income. The table indicates the small size of these additional: while segments A through C all include more than 20% of the households under examination, segments D and E both include less than 6% of the total households. HOT trip rates within segments C through E, formerly the Higher income segment, also differ. The lowest of the segments income-wise, segment C, more closely resembles segments A and B in its rate of exclusive toll lane use: 14.4% versus 14.6% for the Lower income segment and 13.9% for the Medium income segment. At higher incomes, however, those rates increase: Segment D sees an 18.4% rate of HOT-exclusive trips, while Segment E yields a 24% rate of HOT-exclusive trips. GP-exclusive trip taking also decreases with the three sub-segments, going from 48.5% for Segment C to 44.2% in Segment D and then finally to 41.3% in Segment E. Trips by Segment E users have the highest average speed, though no discernible pattern is present in the segment count averages.

Table 102: Expanded 2013 Data Overview – Five Income Segments

	Full Dataset	Segment A	Segment B	Segment C	Segment D	Segment E
Households Analyzed	36,854	10,127	15,588	8,208	1,932	999
% of Households by Income	100%	27.5%	42.3%	22.3%	5.2%	2.7%
Transponders Analyzed	68,325	19,424	28,907	14,610	3,492	1,931
Total Trips Monitored	2,656,430	780,364	1,206,121	527,287	95,262	47,396
HOT-Exclusive Trips	386,370	113,915	167,577	75,949	17,574	11,355
GP-Exclusive Trips	1,337,286	409,743	610,314	255,543	42,107	19,579
Mixed Trips	932,774	256,706	428,230	195,795	35,581	16,462
% of HOT-Exclusive Trips	14.6%	14.6%	13.9%	14.4%	18.4%	24.0%
% of GP-Exclusive Trips	50.3%	52.5%	50.6%	48.5%	44.2%	41.3%
% of Mixed Trips	35.1%	32.9%	35.5%	37.1%	37.4%	34.7%
% of Total Trips by Income		29.4%	45.4%	19.9%	3.6%	1.8%
% of HOT Trips by Income		29.5%	43.4%	19.7%	4.5%	2.9%
% of GP Trips by Income		30.6%	45.6%	19.1%	3.1%	1.5%
% of Mixed Trips by Income		27.5%	45.9%	21.0%	3.8%	1.8%
Average Trip Speed (mph)	53.3	53.0	53.3	53.3	53.4	55.2
Average Segment Count	3.7	3.5	3.7	3.8	3.7	3.7

The next two tables, Table 103 and Table 104, present the results from the five-segment estimation of Model 14b from the previous section. Segments D and E have the highest shares of toll lane alternatives selected, with Segment E exhibiting the highest share of all. The results from Segments A and B are identical to those of the Lower and Medium segments in Table 100, as expected. Many of the estimated coefficients within the previously-singular Higher income segment differ when estimated across the three subsegments examined here. This is first evident in the constant term, for which Segment E exhibits the largest, most positive coefficient. Segment E also yields the smallest coefficient for the toll amount, though the estimator has the same negative sign and is within the same order of magnitude as the other four. Segment E is the only segment to yield a transponderCount coefficient that does not achieve significance at the 95% confidence level; again, all of the estimators for this factor are very close to zero. Within

the household education category, Segment D stands out as having the largest response to this factor. In the household age factor, Segment E has the largest coefficient magnitude.

The last two notable differences appear in the coefficients for the household income and household size factors. Among the household income estimators, Segment D stands out in that it has the only positive coefficient. Only among households in the $150-200k annual income range does additional income increase the probability of toll lane use. Segment E has the negative coefficient of the largest magnitude; among those users, an increase in income yields the largest decrease in toll lane probability across all income segments. Note that this segment still has the largest positive intercept term. The household size estimators also differ within these high income segments: Segment D has an insignificant coefficient that is very close to zero, while Segment E once again has the largest negative estimator.

Among the other factors, the models are more similar than different. All five segments yield negative and significant, but very small, estimators for the square of the average speed difference. The transponderCount coefficients are similarly very small in magnitude; here only the Segment E estimator does not achieve significance at the 95% confidence level. Segment D has the largest response to congested conditions in the general purpose lanes, though its coefficient differs from those of the other segments by a maximum of 0.275. Among the segmentCount, day of week, month of year, and time of day categories, no notable differences are present. The largest discrepancies within these factors occur in the coefficients for trips starting at 9:00 or 9:30 AM: users in Segments D and E have lower probabilities of taking the toll lane at these times, relative to users in the other three income segments.

Table 103: Model 14b with 5 Income Segments - AM Peak

	Segment A	Segment B	Segment C	Segment D	Segment E
Intercept	-1.247*** (t = -8.971)	2.328*** (t = 9.177)	-1.861*** (t = -3.155)	-14.513*** (t = -6.497)	8.638*** (t = 3.672)
$avgSpeed^2$	-0.0003*** (t = -51.375)	-0.0003*** (t = -80.858)	-0.0004*** (t = -58.889)	-0.0004*** (t = -24.007)	-0.0003*** (t = -12.449)
tollAmount	-0.694*** (t = -183.049)	-0.701*** (t = -237.269)	-0.699*** (t = -157.507)	-0.602*** (t = -56.581)	-0.461*** (t = -29.990)
transponderCount	-0.001*** (t = -37.234)	-0.0002*** (t = -5.800)	0.0002*** (t = 5.465)	-0.0004*** (t = -4.058)	-0.0002 (t = -1.340)
HOT: congested50	1.446*** (t = 105.719)	1.482*** (t = 133.393)	1.456*** (t = 86.940)	1.700*** (t = 41.020)	1.425*** (t = 24.350)
HOT: hhEdu	-0.142*** (t = -27.601)	-0.110*** (t = -24.393)	-0.107*** (t = -14.220)	-0.322*** (t = -15.989)	-0.102*** (t = -4.095)
HOT: hhAge	-0.030*** (t = -9.138)	-0.001 (t = -0.479)	-0.094*** (t = -19.132)	-0.104*** (t = -8.207)	-0.515*** (t = -22.753)
HOT:I(hhIncomeDollars)/hhSize)	0 (t = -0.725)	0.00002*** (t = 26.440)	0.00001*** (t = 13.246)	0.00001*** (t = 5.651)	0 (t = -0.662)
HOT: log(hhIncomeDollars)	-0.177*** (t = -11.353)	-0.545*** (t = -22.225)	-0.114** (t = -2.185)	1.058*** (t = 5.561)	-0.548*** (t = -2.803)
HOT: hhSize	-0.056*** (t = -10.499)	0.045*** (t = 11.012)	0.016*** (t = 2.784)	0.01 (t = 0.646)	-0.165*** (t = -7.466)
HOT: segmentCount	0.989*** (t = 199.890)	0.960*** (t = 242.991)	0.989*** (t = 164.322)	0.886*** (t = 62.148)	0.681*** (t = 33.415)
HOT: february	0.310*** (t = 16.187)	0.245*** (t = 16.182)	0.269*** (t = 11.917)	0.163*** (t = 2.997)	0.141* (t = 1.834)
HOT: march	0.240*** (t = 12.896)	0.253*** (t = 17.138)	0.241*** (t = 10.961)	0.244*** (t = 4.566)	0.279*** (t = 3.706)
HOT: april	0.413*** (t = 21.942)	0.398*** (t = 26.508)	0.414*** (t = 18.504)	0.235*** (t = 4.323)	0.335*** (t = 4.336)
HOT: may	0.387*** (t = 20.930)	0.379*** (t = 25.710)	0.429*** (t = 19.505)	0.338*** (t = 6.384)	0.309*** (t = 4.051)
HOT: june	0.255*** (t = 13.170)	0.319*** (t = 20.738)	0.285*** (t = 12.415)	0.321*** (t = 5.689)	0.217*** (t = 2.677)
HOT: july	0.101*** (t = 5.416)	0.198*** (t = 13.368)	0.131*** (t = 5.915)	0.170*** (t = 3.147)	0.225*** (t = 2.925)
HOT: august	0.403*** (t = 21.405)	0.444*** (t = 29.853)	0.442*** (t = 19.892)	0.338*** (t = 6.287)	0.329*** (t = 4.232)
HOT: september	0.586*** (t = 30.504)	0.633*** (t = 41.392)	0.651*** (t = 28.554)	0.585*** (t = 10.615)	0.392*** (t = 4.944)
HOT: october	0.587*** (t = 30.685)	0.653*** (t = 42.934)	0.661*** (t = 29.069)	0.591*** (t = 10.829)	0.485*** (t = 6.107)
HOT: november	0.408*** (t = 20.960)	0.549*** (t = 35.634)	0.529*** (t = 22.769)	0.524*** (t = 9.391)	0.392*** (t = 4.896)
HOT: december	0.172*** (t = 8.884)	0.262*** (t = 17.094)	0.304*** (t = 13.162)	0.353*** (t = 6.365)	0.174** (t = 2.206)
HOT: tuesday	0.290*** (t = 23.706)	0.209*** (t = 21.388)	0.216*** (t = 14.849)	0.264*** (t = 7.512)	0.250*** (t = 4.912)
HOT: wednesday	0.289*** (t = 23.420)	0.203*** (t = 20.588)	0.230*** (t = 15.629)	0.209*** (t = 5.903)	0.238*** (t = 4.620)
HOT: thursday	0.276*** (t = 22.400)	0.211*** (t = 21.490)	0.189*** (t = 12.841)	0.231*** (t = 6.500)	0.203*** (t = 3.954)
HOT: friday	-0.924*** (t = -67.837)	-0.917*** (t = -84.405)	-0.952*** (t = -57.814)	-0.752*** (t = -18.897)	-0.580*** (t = -10.484)
HOT:am630	1.686*** (t = 98.432)	1.579*** (t = 117.295)	1.613*** (t = 77.406)	1.472*** (t = 28.412)	1.146*** (t = 16.707)
HOT: am700	2.061*** (t = 111.143)	1.898*** (t = 129.459)	2.021*** (t = 90.437)	2.062*** (t = 37.491)	1.405*** (t = 17.963)
HOT: am730	2.190*** (t = 115.371)	1.906*** (t = 127.507)	2.126*** (t = 93.174)	1.887*** (t = 34.126)	1.556*** (t = 19.739)
HOT:am800	2.071*** (t = 110.020)	1.805*** (t = 122.752)	1.848*** (t = 81.315)	1.600*** (t = 29.203)	1.622*** (t = 20.703)
HOT:am830	1.843*** (t = 99.346)	1.523*** (t = 104.676)	1.554*** (t = 69.885)	1.545*** (t = 29.073)	0.843*** (t = 11.172)
HOT:am900	1.237*** (t = 67.561)	1.033*** (t = 73.026)	1.129*** (t = 51.802)	0.866*** (t = 16.140)	0.672*** (t = 9.167)
HOT:am930	0.375*** (t = 19.969)	0.283*** (t = 19.350)	0.316*** (t = 13.897)	0.051 (t = 0.921)	0.052 (t = 0.727)
HOT Share	0.5068	0.5202	0.5366	0.6098	0.6734
Observations	342,209	533,623	238,565	41,962	20,655
R^2	0.171	0.166	0.17	0.163	0.128
Log Likelihood	-196,702.70	-308,201.30	-136,646.30	-23,500.00	-11,380.52

The results from the five-segment models for the afternoon peak period of 2013 are shown below in Table 104. As in the AM peak models, the share of toll lane trips increases with the income segments; Segment E once again has the highest share. Unlike the morning models, the goodness of fit measures also increase with income: Segment E has the highest pseudo-R^2 value as well. The remaining differences begin again with the intercept term. In this case, Segment E once again has the constant with the largest magnitude, though in this case it is negative. Segment E also exhibits the lowest sensitivity to the toll amount of a given trip. This relationship is also true for the household education measure: while all of the estimators are negative and significant, the Segment E coefficient is the smallest in magnitude.

The household age coefficients reveal a pattern of decreasing impact on toll lane probability across the five segments: the coefficient for Segment A is small and positive, and the remaining coefficients decrease through Segment E, which has the lowest coefficient. Household income does not yield such a neat pattern: Segments A through C have negative estimators, while those of Segments D and E are positive. Segment E in particular has the largest coefficient magnitude; the remainders are all an order of magnitude smaller. Segment C is unique in the household size category: it has the only positive estimator. Finally, time of day appears to affect lane choice decisions among Segment E users the least: of the seven different intervals, only three yield significant coefficients for those users. Once again, the corridor conditions factors (square of average speed difference, transponder count in both lane types, and congested condition dummy variables) yield very similar results across all segments.

Table 104: Model 14b with 5 Income Segments - PM Peak

	Segment A	Segment B	Segment C	Segment D	Segment E
Intercept	-1.142*** (t = -8.035)	-1.128*** (t = -4.261)	1.521** (t = 2.399)	-1.561 (t = -0.693)	-19.060*** (t = -7.682)
avgSpeed^2	-0.0002*** (t = -34.684)	-0.0002*** (t = -29.407)	-0.0001*** (t = -14.110)	-0.0001*** (t = -4.836)	0.00001 (t = 0.406)
tollAmount	-0.431*** (t = -98.171)	-0.467*** (t = -136.133)	-0.501*** (t = -93.058)	-0.435*** (t = -34.671)	-0.266*** (t = -13.888)
transponderCount	0.003*** (t = 61.096)	0.004*** (t = 99.784)	0.004*** (t = 71.745)	0.004*** (t = 30.650)	0.004*** (t = 17.909)
HOT: congested40	1.367*** (t = 135.420)	1.344*** (t = 165.121)	1.424*** (t = 111.760)	1.370*** (t = 45.825)	1.273*** (t = 28.346)
HOT: hhEdu	-0.200*** (t = -38.167)	-0.192*** (t = -40.699)	-0.308*** (t = -38.326)	-0.326*** (t = -16.261)	-0.084*** (t = -3.215)
HOT: hhAge	0.012*** (t = 3.759)	0.001 (t = 0.512)	-0.072*** (t = -13.974)	-0.074*** (t = -5.772)	-0.184*** (t = -8.291)
HOT:I(hhIncomeDollars)/hhSize)	0.00000** (t = 1.985)	0.00000*** (t = 4.717)	0.00001*** (t = 13.917)	-0.00002*** (t = -9.602)	0 (t = 1.013)
HOT: log(hhIncomeDollars)	-0.111*** (t = -7.005)	-0.112*** (t = -4.359)	-0.340*** (t = -6.065)	0.126 (t = 0.661)	1.407*** (t = 6.873)
HOT: hhSize	-0.013** (t = -2.462)	-0.020*** (t = -4.617)	0.075*** (t = 12.201)	-0.134*** (t = -8.794)	-0.161*** (t = -6.961)
HOT: segmentCount	1.068*** (t = 218.616)	1.079*** (t = 272.467)	1.152*** (t = 183.575)	1.193*** (t = 80.131)	1.146*** (t = 53.055)
HOT: february	-0.064*** (t = -3.265)	-0.067*** (t = -4.286)	-0.094*** (t = -3.812)	-0.143** (t = -2.493)	-0.021 (t = -0.253)
HOT: march	-0.042** (t = -2.160)	-0.054*** (t = -3.484)	-0.084*** (t = -3.412)	-0.133** (t = -2.343)	-0.095 (t = -1.132)
HOT: april	-0.250*** (t = -12.897)	-0.272*** (t = -17.353)	-0.286*** (t = -11.668)	-0.443*** (t = -7.679)	-0.492*** (t = -5.863)
HOT: may	-0.376*** (t = -19.675)	-0.356*** (t = -23.082)	-0.393*** (t = -16.355)	-0.446*** (t = -7.876)	-0.487*** (t = -5.854)
HOT: june	-0.395*** (t = -20.012)	-0.333*** (t = -20.980)	-0.401*** (t = -16.174)	-0.372*** (t = -6.233)	-0.431*** (t = -4.939)
HOT: july	-0.366*** (t = -18.800)	-0.314*** (t = -19.978)	-0.430*** (t = -17.573)	-0.403*** (t = -6.924)	-0.578*** (t = -6.661)
HOT: august	-0.184*** (t = -9.500)	-0.097*** (t = -6.282)	-0.187*** (t = -7.762)	-0.228*** (t = -4.038)	-0.471*** (t = -5.589)
HOT: september	0.139*** (t = 7.012)	0.328*** (t = 20.554)	0.281*** (t = 11.349)	0.155*** (t = 2.687)	-0.273*** (t = -3.159)
HOT: october	0.178*** (t = 9.098)	0.375*** (t = 23.748)	0.311*** (t = 12.590)	0.188*** (t = 3.300)	-0.273*** (t = -3.223)
HOT: november	0.092*** (t = 4.621)	0.325*** (t = 20.300)	0.303*** (t = 12.133)	0.065 (t = 1.125)	-0.358*** (t = -4.167)
HOT: december	0.118*** (t = 5.971)	0.266*** (t = 16.593)	0.270*** (t = 10.795)	0.137** (t = 2.336)	-0.253*** (t = -2.963)
HOT: tuesday	0.024* (t = 1.954)	0.012 (t = 1.198)	0.017 (t = 1.107)	-0.042 (t = -1.158)	0.038 (t = 0.714)
HOT: wednesday	0.049*** (t = 3.960)	0.022** (t = 2.221)	0.008 (t = 0.536)	-0.023 (t = -0.633)	-0.012 (t = -0.222)
HOT: thursday	-0.069*** (t = -5.316)	-0.117*** (t = -11.095)	-0.129*** (t = -7.823)	-0.258*** (t = -6.688)	-0.172*** (t = -3.016)
HOT: friday	-0.031** (t = -2.261)	-0.120*** (t = -10.640)	-0.116*** (t = -6.594)	-0.169*** (t = -4.062)	0.082 (t = 1.341)
HOT:pm1530	-0.001 (t = -0.037)	-0.058*** (t = -3.861)	-0.114*** (t = -5.101)	-0.285*** (t = -5.650)	0.047 (t = 0.649)
HOT:pm1600	-0.108*** (t = -5.874)	-0.156*** (t = -10.379)	-0.128*** (t = -5.605)	-0.186*** (t = -3.565)	-0.092 (t = -1.238)
HOT:pm1630	-0.068*** (t = -3.597)	-0.037** (t = -2.429)	0.172*** (t = 7.383)	-0.200*** (t = -3.811)	0.039 (t = 0.507)
HOT:pm1700	0.019 (t = 0.984)	0.088*** (t = 5.728)	0.227*** (t = 9.693)	-0.235*** (t = -4.430)	0.075 (t = 0.975)
HOT:pm1730	0.099*** (t = 5.313)	0.229*** (t = 15.018)	0.346*** (t = 14.965)	0.067 (t = 1.264)	0.606*** (t = 7.896)
HOT:pm1800	0.288*** (t = 15.807)	0.418*** (t = 28.278)	0.554*** (t = 24.439)	0.337*** (t = 6.512)	0.802*** (t = 11.164)
HOT:pm1830	0.241*** (t = 13.371)	0.344*** (t = 23.759)	0.517*** (t = 23.342)	0.207*** (t = 4.080)	0.802*** (t = 11.588)
HOT Share	0.5223	0.5389	0.5633	0.5827	0.5958
Observations	349,783	544,660	235,868	43,494	21,194
R2	0.199	0.213	0.243	0.244	0.277
Log Likelihood	-193,922.40	-295,646.70	-122,360.80	-22,349.61	-10,332.79

Mixed Logit Models

The mixed logit framework used in the following models addresses some of the issues with the standard binary logit models that were previously discussed. Foremost among these is the issue of serial correlation that arises when estimating models with panel data, as the author is doing here. The random parameter estimation of the mixed logit method also allows for better understanding of the range of responses to the model's independent variables (Train, 2002). This section presents the results of the previously designed models estimated with the mixed logit framework to address these issues and reduce the model bias that results from the standard models.

The first of these models is presented in Table 105 and Table 106. These models use the previous Model 14b as the basis for their design, and also separate the data into three income segments and AM/PM peak segments. Additionally, this first pair of mixed logit models sets the tollAmount coefficient to a random parameter with a normal distribution. The tollAmount coefficient reported in the tables represents the mean of that distribution, while the 'tollAmount Standard Deviation' rows report the standard deviation of that distribution. The normally-distributed tollAmount random parameters have standard deviations ranging from 2.038 to 2.244; in all three segments, that standard deviation value achieves significance at the 99% confidence level. This supports the hypothesis that the user responses to toll levels vary within income segments, and include both positive and negative responses.

Like the models in the previous section, the results presented here illustrate similarities and differences across the income segments. As before, the operational characteristics (square of average speed difference, count of transponders, congested

conditions) yield coefficients that are similar in magnitude, sign, and significance for each model. The toll amount estimators indicate that the Lower income segment has the highest sensitivity to toll levels, other factors being equal. The remaining model differences reside primarily in the demographic characteristics: household age, education, income, and size. The lower income segment has the only education coefficient that is not negative and significant. The household age factor is negative and significant at the 95% confidence level in both the Lower and Higher segments; the Medium segment factor is the only positive one. Goodness of fit results as indicated by McFadden's pseudo-R^2 value indicate that the mixed logit framework substantially improves model fit relative to the standard binary logit method.

Table 105: Mixed Logit Model 1a with 3 Income Segments – AM Peak

	AM Peak – Lower	AM Peak – Medium	AM Peak – Higher
Intercept	-4.475*** (t = -19.118)	-3.414*** (t = -7.482)	-8.048*** (t = -13.787)
avgSpeed^2	-0.0004*** (t = -54.510)	-0.001*** (t = -75.295)	-0.001*** (t = -60.901)
tollAmount	-0.816*** (t = -124.283)	-0.752*** (t = -143.762)	-0.632*** (t = -88.396)
transponderCount	0.001*** (t = 16.756)	0.002*** (t = 53.459)	0.002*** (t = 35.873)
HOT: congested50	1.935*** (t = 84.873)	1.971*** (t = 102.604)	2.050*** (t = 78.463)
HOT: hhEdu	0.030*** (t = 3.414)	-0.034*** (t = -4.300)	-0.104*** (t = -8.715)
HOT: hhAge	-0.028*** (t = -5.212)	0.028*** (t = 5.813)	-0.081*** (t = -10.142)
HOT:I(hhIncomeDollars)/hhSize)	-0.00000* (t = -1.780)	0.00000*** (t = 3.104)	-0.00000* (t = -1.755)
HOT: log(hhIncomeDollars)	0.118*** (t = 4.484)	0.025 (t = 0.572)	0.491*** (t = 9.085)
HOT: hhSize	-0.041*** (t = -4.593)	-0.016** (t = -2.187)	-0.041*** (t = -4.783)
HOT: segmentCount	1.030*** (t = 142.041)	1.037*** (t = 176.218)	1.086*** (t = 130.783)
HOT: february	0.456*** (t = 14.371)	0.421*** (t = 16.209)	0.463*** (t = 13.361)
HOT: march	0.465*** (t = 14.864)	0.496*** (t = 19.154)	0.495*** (t = 14.416)
HOT: april	0.675*** (t = 21.324)	0.661*** (t = 24.948)	0.693*** (t = 19.830)
HOT: may	0.526*** (t = 17.499)	0.507*** (t = 20.235)	0.652*** (t = 19.535)
HOT: june	0.535*** (t = 16.642)	0.525*** (t = 19.435)	0.642*** (t = 18.104)
HOT: july	0.309*** (t = 10.249)	0.333*** (t = 13.354)	0.412*** (t = 12.180)
HOT: august	0.627*** (t = 19.864)	0.600*** (t = 23.144)	0.687*** (t = 19.847)
HOT: september	0.704*** (t = 22.223)	0.646*** (t = 24.841)	0.745*** (t = 20.833)
HOT: october	0.772*** (t = 24.511)	0.711*** (t = 27.397)	0.802*** (t = 22.481)
HOT: november	0.459*** (t = 14.981)	0.485*** (t = 19.155)	0.533*** (t = 15.420)
HOT: december	0.050* (t = 1.684)	0.069*** (t = 2.807)	0.163*** (t = 4.943)
HOT: tuesday	0.389*** (t = 18.367)	0.330*** (t = 18.753)	0.331*** (t = 13.913)
HOT: wednesday	0.480*** (t = 22.338)	0.369*** (t = 20.558)	0.402*** (t = 16.639)
HOT: thursday	0.411*** (t = 19.509)	0.375*** (t = 21.351)	0.367*** (t = 15.430)
HOT: friday	-0.798*** (t = -39.095)	-0.808*** (t = -47.771)	-0.890*** (t = -38.737)
HOT:am630	1.762*** (t = 61.354)	1.684*** (t = 71.363)	1.822*** (t = 53.978)
HOT: am700	1.932*** (t = 65.404)	1.861*** (t = 76.617)	2.154*** (t = 63.912)
HOT: am730	1.908*** (t = 64.257)	1.652*** (t = 68.969)	1.908*** (t = 56.733)
HOT:am800	1.460*** (t = 49.912)	1.258*** (t = 53.100)	1.369*** (t = 40.916)
HOT:am830	1.017*** (t = 35.443)	0.835*** (t = 35.704)	0.867*** (t = 26.315)
HOT:am900	0.375*** (t = 13.585)	0.138*** (t = 6.286)	0.203*** (t = 6.539)
HOT:am930	-0.482*** (t = -18.311)	-0.590*** (t = -28.349)	-0.580*** (t = -19.490)
tollAmount Standard Deviation	2.241*** (t = 242.029)	2.244*** (t = 305.948)	2.038*** (t = 223.190)
HOT Share	0.507	0.520	0.556
Observations	342,209	533,623	301,182
R^2	0.599	0.619	0.625
Log Likelihood	-95,048.05	-140,855.00	-77,501.39

*p<0.1; **p<0.05; ***p<0.01

Figure 148 presents the parameter distributions for the tollAmount variable for the AM peak period three-segment models. The three curves are similar in shape and location; this is especially true of the Lower and Medium income segments. Also notable is the proximity of the mean to the zero value; as a result, all three income segments include sizeable portions of their spectrum in both the positive and negative regions. This suggests that the response to toll levels is not a simple, constant positive or negative value, but rather varies along a spectrum that includes both types of responses.

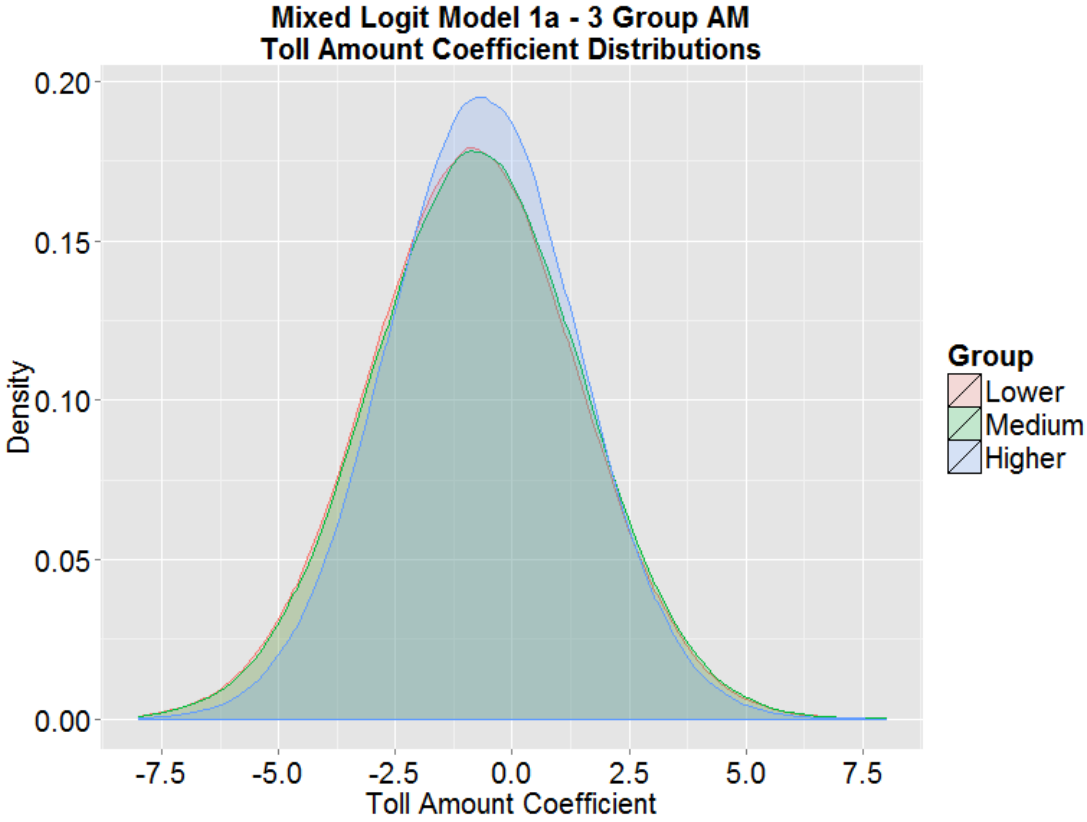


Figure 148: Normal Distributions for Toll Amount Parameter - AM Models

Results for the PM peak period models are presented in Table 106 below. Once again, model fit as represented by McFadden's pseudo-R^2 metric indicates better fits with the mixed logit framework versus the standard binary logit models. These models yield fewer differences across the three income segments; in particular, the toll amount and

household education coefficients are more similar than the morning peak period counterparts. The Higher income segment does yield a toll amount coefficient that is smaller in magnitude than those of the Lower and Medium segments, but the difference is not as stark as it was previously. All three segments fail to achieve significance at the 95% confidence level in their household age estimators, as well as their household income and household size estimators. The magnitudes of these coefficients vary across income groups, but they are close to zero.

Table 106: Mixed Logit Model 1a with 3 Income Segments – PM Peak

	PM Peak – Lower	PM Peak – Medium	PM Peak – Higher
Intercept	-4.171*** (t = -16.276)	-2.753*** (t = -5.849)	-1.488** (t = -2.507)
avgSpeed^2	-0.0002*** (t = -17.749)	-0.0002*** (t = -17.281)	-0.0001*** (t = -11.598)
tollAmount	-0.359*** (t = -43.639)	-0.289*** (t = -44.265)	-0.213*** (t = -23.321)
transponderCount	0.007*** (t = 102.907)	0.008*** (t = 155.630)	0.008*** (t = 116.659)
HOT: congested40	1.709*** (t = 98.157)	1.687*** (t = 116.033)	1.649*** (t = 82.455)
HOT: hhEdu	0.018* (t = 1.922)	-0.031*** (t = -3.869)	-0.016 (t = -1.268)
HOT: hhAge	0.003 (t = 0.517)	-0.001 (t = -0.303)	-0.01 (t = -1.223)
HOT:I(hhIncomeDollars)/hhSize)	0.00001*** (t = 2.931)	0.00001*** (t = 5.600)	0 (t = -0.378)
HOT: log(hhIncomeDollars)	-0.045 (t = -1.576)	-0.190*** (t = -4.165)	-0.286*** (t = -5.201)
HOT: hhSize	0.014 (t = 1.409)	0.034*** (t = 4.650)	0.019** (t = 2.086)
HOT: segmentCount	1.511*** (t = 183.759)	1.587*** (t = 234.261)	1.659*** (t = 174.215)
HOT: february	-0.100*** (t = -3.182)	-0.086*** (t = -3.278)	-0.152*** (t = -4.259)
HOT: march	-0.088*** (t = -2.792)	-0.087*** (t = -3.315)	-0.126*** (t = -3.540)
HOT: april	-0.276*** (t = -8.700)	-0.284*** (t = -10.777)	-0.383*** (t = -10.565)
HOT: may	-0.472*** (t = -15.123)	-0.419*** (t = -16.190)	-0.480*** (t = -13.650)
HOT: june	-0.471*** (t = -14.847)	-0.436*** (t = -16.323)	-0.518*** (t = -14.023)
HOT: july	-0.467*** (t = -15.092)	-0.419*** (t = -16.157)	-0.528*** (t = -14.754)
HOT: august	-0.213*** (t = -6.594)	-0.125*** (t = -4.648)	-0.249*** (t = -6.793)
HOT: september	0.236*** (t = 7.098)	0.426*** (t = 15.557)	0.224*** (t = 6.009)
HOT: october	0.310*** (t = 9.294)	0.532*** (t = 19.224)	0.303*** (t = 7.946)
HOT: november	0.176*** (t = 5.343)	0.386*** (t = 14.092)	0.247*** (t = 6.629)
HOT: december	0.091*** (t = 2.810)	0.266*** (t = 9.757)	0.217*** (t = 5.793)
HOT: tuesday	-0.012 (t = -0.578)	0.011 (t = 0.613)	-0.041* (t = -1.716)
HOT: wednesday	0.02 (t = 0.941)	0.016 (t = 0.907)	-0.008 (t = -0.321)
HOT: thursday	-0.131*** (t = -6.023)	-0.145*** (t = -7.953)	-0.198*** (t = -7.985)
HOT: friday	-0.072*** (t = -3.213)	-0.144*** (t = -7.710)	-0.099*** (t = -3.827)
HOT:pm1530	-0.088*** (t = -3.140)	-0.120*** (t = -5.153)	-0.193*** (t = -6.303)
HOT:pm1600	-0.160*** (t = -5.519)	-0.195*** (t = -8.237)	-0.127*** (t = -4.018)
HOT:pm1630	0.036 (t = 1.237)	0.038 (t = 1.585)	0.161*** (t = 5.042)
HOT:pm1700	0.214*** (t = 7.193)	0.294*** (t = 12.057)	0.361*** (t = 10.949)
HOT:pm1730	0.400*** (t = 13.457)	0.509*** (t = 20.793)	0.626*** (t = 19.120)
HOT:pm1800	0.646*** (t = 22.531)	0.652*** (t = 27.309)	0.755*** (t = 23.336)
HOT:pm1830	0.521*** (t = 18.764)	0.550*** (t = 23.691)	0.598*** (t = 19.306)
tollAmount Standard Deviation	3.182*** (t = 255.627)	3.275*** (t = 313.744)	3.060*** (t = 226.386)
HOT Share	0.522	0.539	0.568
Observations	348,894	544,660	300,556
R^2	0.601	0.625	0.625
Log Likelihood	-96,320.97	-140,890.70	-77,082.36

*p<0.1; **p<0.05; ***p<0.01

Figure 149 presents the PM Peak parameter distributions for the tollAmount variable. The Medium income segment differs more from the Lower and Higher segments here in its larger standard deviation: the resulting distribution is wider and less

peaked. The difference between the segments is less pronounced here than in the morning peak. Like the AM peak period results, the means of these distributions are close to zero and the range of responses includes both positive and negative values.

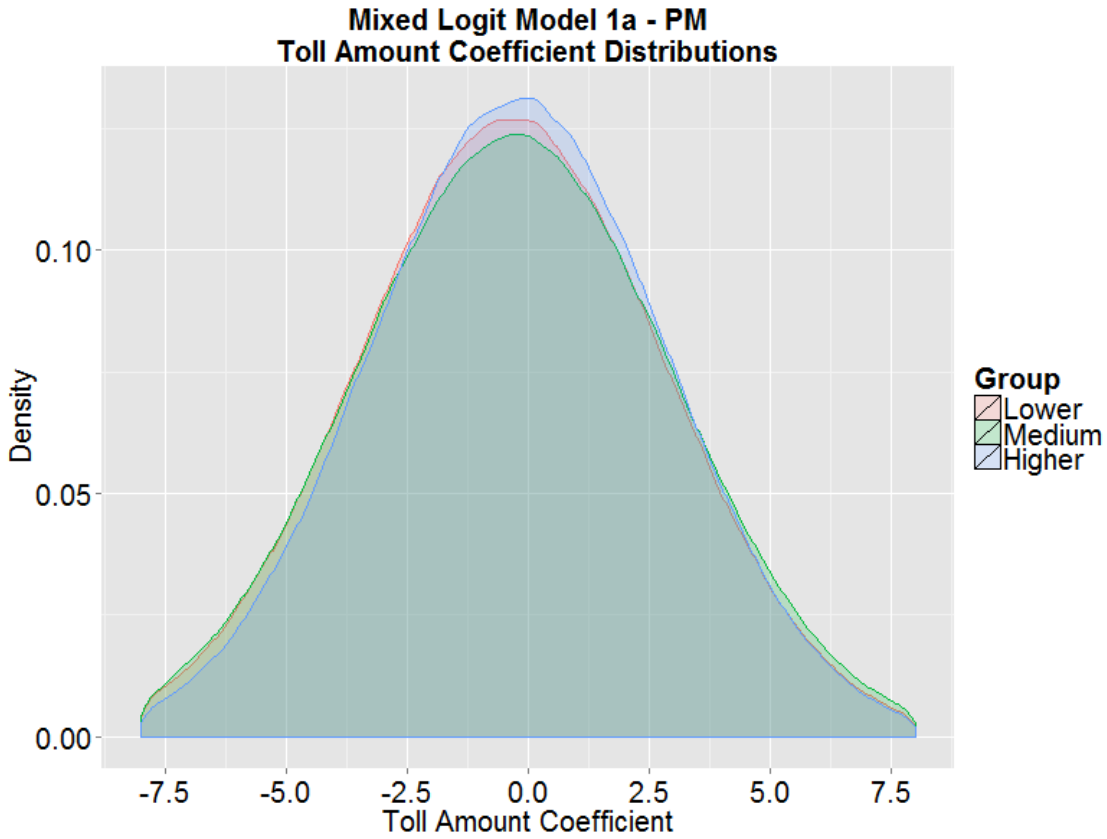


Figure 149: Normal Distributions for Toll Amount Parameter - PM Models

The following pair of models is identical in design to the first mixed logit models with one change: the toll amount parameter is now estimated with a log-normal distribution rather than a normal distribution. The toll amount itself is multiplied by negative one to make the value negative, so that the log-normal distribution is positive.

Among the AM peak period models, the largest difference is in the coefficients of the (now negative) toll amount term. Whereas previously all three segments had negative coefficients that achieved significance at the 95% confidence level (and represented means of the normal distribution), here only the Medium segment estimator (and mean of

the log-normal distribution) achieves significance. The other two values are actually negative, though again their t-statistics are sufficiently low that the author cannot reject the null hypothesis that they are equal to zero. The lack of uniform significance in the toll amount coefficients indicates that the model with the normal distribution on the toll amount is a better choice.

Table 107: Mixed Logit Model 1b with 3 Income Segments – AM Peak

	AM Peak – Lower	AM Peak – Medium	AM Peak – Higher
Intercept	-1.774 (t = -1.118)	8.606** (t = 2.264)	-13.959*** (t = -3.901)
avgSpeed^2	-0.0004*** (t = -6.523)	-0.0005*** (t = -8.292)	-0.001*** (t = -9.995)
-1*tollAmount	-0.012 (t = -0.243)	0.193*** (t = 4.049)	-0.05 (t = -1.018)
transponderCount	-0.001* (t = -1.827)	0.001*** (t = 3.182)	0.001 (t = 1.347)
HOT: congested50	2.396*** (t = 15.244)	2.388*** (t = 13.964)	2.602*** (t = 15.919)
HOT: hhEdu	-0.019 (t = -0.340)	-0.259*** (t = -3.797)	-0.286*** (t = -3.965)
HOT: hhAge	-0.162*** (t = -4.749)	-0.043 (t = -1.047)	-0.135*** (t = -2.864)
HOT:I(hhIncomeDollars)/hhSize)	0.00001 (t = 0.784)	0.0001*** (t = 5.324)	0.00001 (t = 0.776)
HOT: log(hhIncomeDollars)	-0.134 (t = -0.748)	-1.131*** (t = -3.047)	1.082*** (t = 3.261)
HOT: hhSize	0.099 (t = 1.557)	0.192*** (t = 3.053)	-0.089* (t = -1.646)
HOT: segmentCount	1.100*** (t = 24.355)	1.227*** (t = 24.085)	1.202*** (t = 23.196)
HOT: february	0.415** (t = 1.996)	0.377* (t = 1.718)	0.465** (t = 2.194)
HOT: march	0.433** (t = 2.161)	0.458** (t = 2.173)	0.393* (t = 1.938)
HOT: april	0.460** (t = 2.202)	0.450** (t = 2.122)	0.785*** (t = 3.565)
HOT: may	0.3 (t = 1.543)	0.271 (t = 1.259)	0.623*** (t = 2.902)
HOT: june	0.295 (t = 1.540)	0.373* (t = 1.679)	0.961*** (t = 4.557)
HOT: july	-0.091 (t = -0.489)	0.1 (t = 0.495)	0.498** (t = 2.494)
HOT: august	0.476** (t = 2.341)	0.262 (t = 1.195)	0.869*** (t = 4.318)
HOT: september	0.605*** (t = 3.103)	0.652*** (t = 3.006)	0.817*** (t = 3.737)
HOT: october	0.304 (t = 1.442)	0.855*** (t = 3.951)	0.977*** (t = 4.569)
HOT: november	0.352* (t = 1.786)	0.3 (t = 1.452)	0.420* (t = 1.914)
HOT: december	-0.113 (t = -0.603)	-0.262 (t = -1.388)	0.269 (t = 1.324)
HOT: tuesday	0.486*** (t = 3.337)	0.091 (t = 0.607)	0.344** (t = 2.261)
HOT: wednesday	0.706*** (t = 4.775)	0.459*** (t = 3.011)	0.139 (t = 0.928)
HOT: thursday	0.711*** (t = 5.080)	0.213 (t = 1.404)	0.329** (t = 2.171)
HOT: friday	-0.769*** (t = -6.152)	-1.256*** (t = -8.906)	-1.140*** (t = -8.589)
HOT:am630	1.922*** (t = 10.271)	2.543*** (t = 12.372)	2.048*** (t = 10.356)
HOT: am700	2.390*** (t = 12.587)	2.826*** (t = 13.167)	2.759*** (t = 13.459)
HOT: am730	2.324*** (t = 12.138)	2.591*** (t = 12.820)	2.476*** (t = 12.746)
HOT:am800	1.810*** (t = 10.263)	2.381*** (t = 11.840)	2.181*** (t = 11.500)
HOT:am830	1.573*** (t = 8.687)	1.917*** (t = 9.956)	1.538*** (t = 8.158)
HOT:am900	0.901*** (t = 5.197)	0.949*** (t = 5.624)	0.650*** (t = 3.866)
HOT:am930	-0.138 (t = -0.883)	-0.197 (t = -1.274)	-0.492*** (t = -3.109)
-1*tollAmount Standard Deviation	1.570*** (t = 21.981)	1.212*** (t = 23.436)	1.358*** (t = 24.030)
HOT Share	0.506	0.521	0.549
Observations	10,000	10,000	10,000
R^2	0.333	0.31	0.372
Log Likelihood	-4,621.92	-4,777.70	-4,320.12

*p<0.1; **p<0.05; ***p<0.01

The resulting distributions of the log-normal toll amount coefficient are presented in Figure 150. The Lower and Higher income segments resemble each other more closely than the Medium income segment. While the Medium segment was the only model with a significant toll amount coefficient (and thus log-normal distributional

mean), all three distributions had positive, statistically significant standard deviation values. As in the previous model, it is evident that representing response to toll amounts is better handled with a range of values rather than a single estimated coefficient.

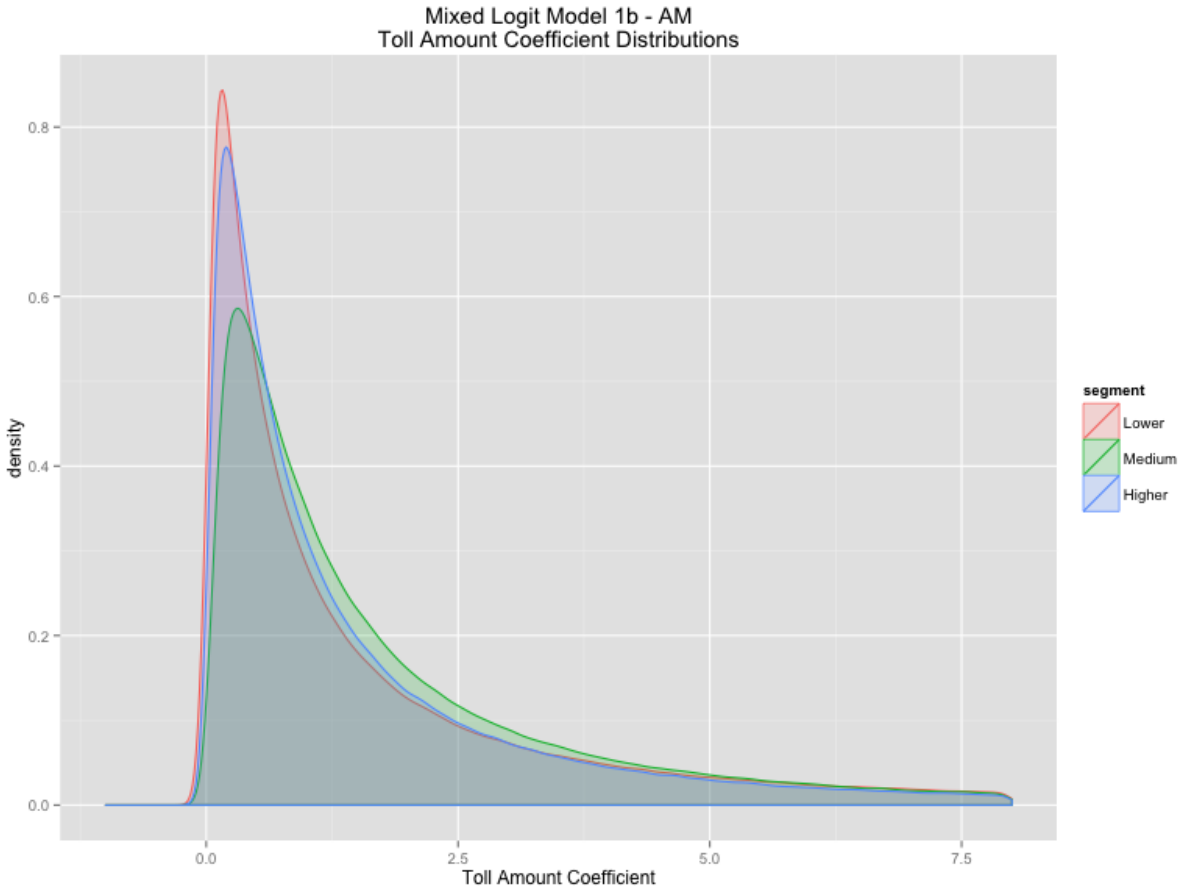


Figure 150: Log-normal Distributions for Toll Amount Parameter - AM Models

Table 108 presents the results from the PM peak period models with negative toll amount values and log-normal toll amount coefficient distributions. Unlike the AM models, here the coefficient estimates (and distribution means) all achieve statistical significance at the 95% confidence level.

Table 108: Mixed Logit Model 1b with 3 Income Segments – PM Peak

	PM Peak – Lower	PM Peak – Medium	PM Peak – Higher
Intercept	-1.925 (t = -1.183)	-7.975** (t = -2.303)	-3.521 (t = -1.023)
$avgSpeed^2$	-0.0003*** (t = -4.337)	-0.0002*** (t = -2.741)	-0.0002*** (t = -2.843)
-1*tollAmount	-0.264*** (t = -3.412)	-0.371*** (t = -4.742)	-0.362*** (t = -4.702)
transponderCount	0.005*** (t = 9.436)	0.006*** (t = 14.233)	0.005*** (t = 11.331)
HOT: congested40	2.417*** (t = 20.233)	2.453*** (t = 20.397)	2.376*** (t = 19.216)
HOT: hhEdu	-0.210*** (t = -3.582)	-0.243*** (t = -4.052)	-0.239*** (t = -3.516)
HOT: hhAge	-0.102*** (t = -3.038)	-0.003 (t = -0.093)	0.041 (t = 0.851)
HOT:I(hhIncomeDollars)/hhSize)	0 (t = 0.136)	-0.00002** (t = -1.978)	0 (t = 0.105)
HOT: log(hhIncomeDollars)	-0.065 (t = -0.360)	0.533 (t = 1.583)	-0.014 (t = -0.044)
HOT: hhSize	-0.033 (t = -0.549)	-0.143** (t = -2.571)	-0.03 (t = -0.555)
HOT: segmentCount	1.643*** (t = 29.452)	1.519*** (t = 29.416)	1.687*** (t = 30.183)
HOT: february	0.115 (t = 0.602)	0.208 (t = 1.146)	0.08 (t = 0.398)
HOT: march	0.395** (t = 2.004)	-0.199 (t = -1.102)	0.159 (t = 0.848)
HOT: april	0.013 (t = 0.063)	-0.101 (t = -0.545)	0.048 (t = 0.235)
HOT: may	-0.166 (t = -0.880)	-0.285 (t = -1.579)	-0.307 (t = -1.590)
HOT: june	-0.269 (t = -1.305)	-0.202 (t = -1.090)	-0.279 (t = -1.347)
HOT: july	-0.374** (t = -2.005)	-0.367** (t = -2.029)	-0.499** (t = -2.487)
HOT: august	0.17 (t = 0.845)	-0.027 (t = -0.139)	0.016 (t = 0.079)
HOT: september	0.676*** (t = 3.308)	0.496** (t = 2.538)	0.209 (t = 1.001)
HOT: october	0.833*** (t = 3.905)	0.469** (t = 2.284)	0.345* (t = 1.682)
HOT: november	0.366* (t = 1.788)	0.469** (t = 2.313)	0.590*** (t = 2.643)
HOT: december	0.466** (t = 2.282)	0.25 (t = 1.166)	0.383* (t = 1.764)
HOT: tuesday	0.171 (t = 1.377)	0.194 (t = 1.575)	0.011 (t = 0.089)
HOT: wednesday	0.204 (t = 1.564)	0.085 (t = 0.678)	0.079 (t = 0.587)
HOT: thursday	-0.055 (t = -0.408)	0.149 (t = 1.135)	-0.029 (t = -0.208)
HOT: friday	0.024 (t = 0.168)	0.075 (t = 0.532)	0.035 (t = 0.238)
HOT:pm1530	-0.06 (t = -0.339)	0.012 (t = 0.074)	0.211 (t = 1.242)
HOT:pm1600	-0.279 (t = -1.517)	0.071 (t = 0.417)	0.353* (t = 1.927)
HOT:pm1630	-0.092 (t = -0.479)	0.252 (t = 1.441)	0.848*** (t = 4.535)
HOT:pm1700	0.098 (t = 0.508)	0.299* (t = 1.683)	0.894*** (t = 4.793)
HOT:pm1730	0.11 (t = 0.577)	0.457** (t = 2.504)	1.035*** (t = 5.625)
HOT:pm1800	0.458*** (t = 2.652)	0.576*** (t = 3.492)	1.154*** (t = 6.565)
HOT:pm1830	0.244 (t = 1.456)	0.556*** (t = 3.470)	0.963*** (t = 5.683)
-1*tollAmount Standard Deviation	2.012*** (t = 18.671)	1.861*** (t = 19.092)	1.930*** (t = 19.161)
HOT Share	0.530	0.536	0.569
Observations	10,000	10,000	10,000
R^2	0.338	0.315	0.373
Log Likelihood	-4,577.86	-4,730.18	-4,287.08

*p<0.1; **p<0.05; ***p<0.01

Figure 151 illustrates the log-normal distributions for the negative toll amount parameters in the PM peak mixed logit models. Compared to the AM peak distributions, the results are much more uniform, with little to distinguish them from each other. While the normally distributed toll amount coefficients were similar, the log-normal coefficient distributions are even more so.

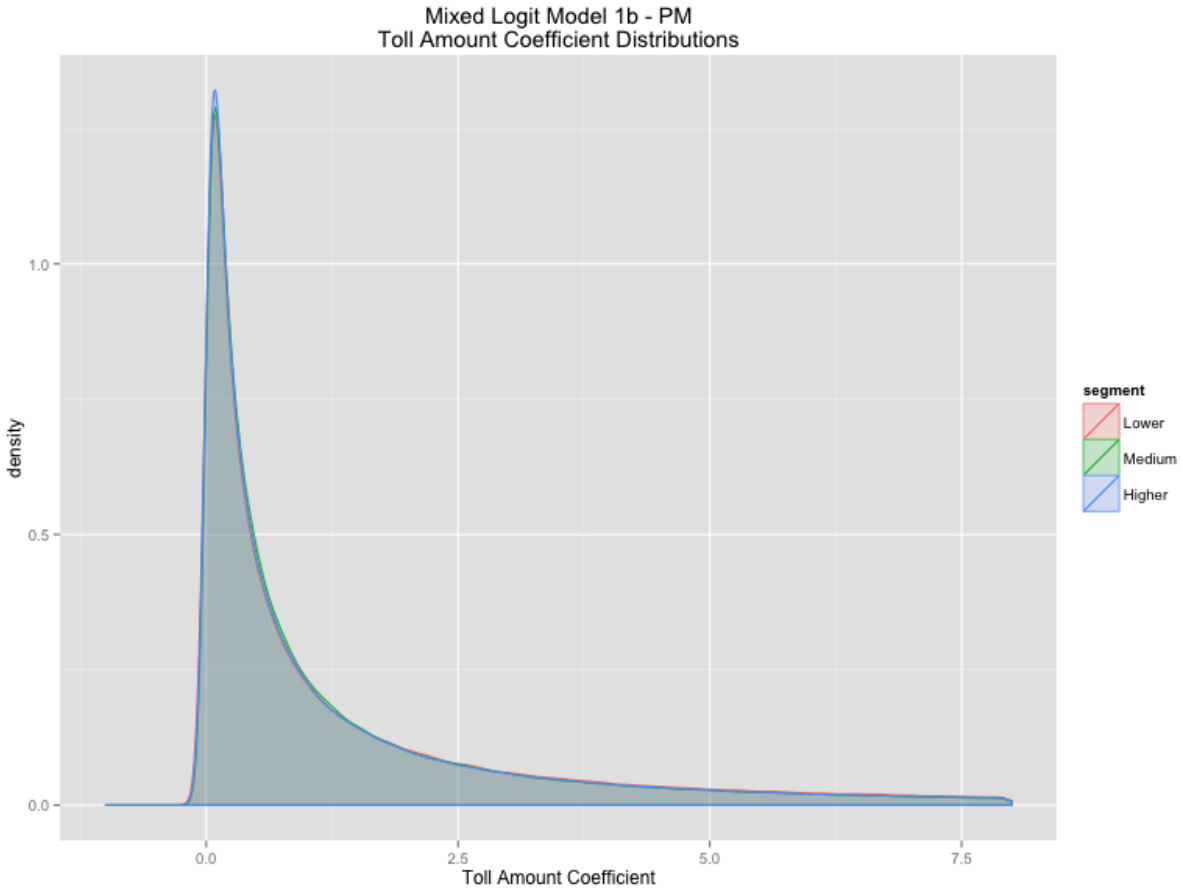


Figure 151: Log-normal Distributions for Toll Amount Parameter - PM Models

The final pair of three-segment mixed logit models is presented below in Table 109 and Table 110. These models randomize the parameters estimated for the household income, rather than the toll amoung, using the normal distribution. The motivation for this was to compare the range of responses to household income across the different segments to investigate whether any substantial differences were present. Table 109

displays the results from the AM peak period models. As before, the majority of the model results remain similar in sign and magnitude. The household income results themselves follow the pattern established in previous models, in which the Lower income coefficient is close to zero but does not achieve significance at the 95% confidence level. The Medium income coefficient is negative and significant, while the Higher income coefficient is positive and significant, with the largest magnitude. The estimated standard deviations of the three normal distributions are significant at the 99% confidence level, indicating the appropriateness of representing household income response as a range rather than as a fixed value.

Table 109: Mixed Logit Model 2 with 3 Income Segments – AM Peak

	AM Peak – Lower	AM Peak – Medium	AM Peak – Higher
Intercept	-4.579*** (t = -2.621)	11.029** (t = 2.490)	-41.208*** (t = -9.400)
avgSpeed^2	-0.001*** (t = -10.658)	-0.001*** (t = -9.995)	-0.001*** (t = -11.326)
tollAmount	-1.402*** (t = -21.537)	-1.530*** (t = -23.615)	-1.392*** (t = -24.002)
transponderCount	0.001 (t = 1.292)	0.001*** (t = 2.687)	0.002*** (t = 5.176)
HOT: congested50	3.241*** (t = 16.130)	3.093*** (t = 14.739)	2.743*** (t = 14.559)
HOT: hhEdu	-0.254*** (t = -3.716)	-0.512*** (t = -6.319)	-0.399*** (t = -4.659)
HOT: hhAge	-0.046 (t = -1.109)	0.026 (t = 0.526)	-0.220*** (t = -3.889)
HOT:I(hhIncomeDollars)/hhSize)	-0.00003 (t = -1.250)	0.00005*** (t = 3.852)	0 (t = -0.438)
HOT: log(hhIncomeDollars)	-0.101 (t = -0.516)	-1.599*** (t = -3.713)	3.242*** (t = 8.062)
HOT: hhSize	-0.259*** (t = -3.695)	0.078 (t = 1.153)	-0.199*** (t = -3.232)
HOT: segmentCount	2.261*** (t = 25.356)	2.135*** (t = 24.408)	2.178*** (t = 28.142)
HOT: february	0.680*** (t = 2.667)	0.042 (t = 0.161)	0.161 (t = 0.657)
HOT: march	0.679*** (t = 2.829)	0.479** (t = 1.972)	0.12 (t = 0.498)
HOT: april	0.992*** (t = 3.978)	0.828*** (t = 3.052)	0.431* (t = 1.721)
HOT: may	0.514** (t = 2.116)	0.970*** (t = 3.803)	0.605** (t = 2.374)
HOT: june	0.653** (t = 2.573)	0.559** (t = 2.072)	0.414 (t = 1.611)
HOT: july	0.094 (t = 0.397)	0.463* (t = 1.804)	-0.191 (t = -0.765)
HOT: august	0.968*** (t = 3.904)	0.746*** (t = 2.958)	0.832*** (t = 3.247)
HOT: september	1.092*** (t = 4.207)	1.410*** (t = 5.298)	0.894*** (t = 3.568)
HOT: october	0.863*** (t = 3.610)	1.458*** (t = 5.515)	0.970*** (t = 3.931)
HOT: november	0.675*** (t = 2.735)	1.309*** (t = 4.961)	0.688*** (t = 2.693)
HOT: december	-0.098 (t = -0.404)	0.08 (t = 0.314)	0.367 (t = 1.457)
HOT: tuesday	0.499*** (t = 2.921)	0.452*** (t = 2.663)	0.443*** (t = 2.656)
HOT: wednesday	0.642*** (t = 3.752)	0.373** (t = 2.145)	0.391** (t = 2.360)
HOT: thursday	0.461*** (t = 2.669)	0.436** (t = 2.509)	0.355** (t = 2.007)
HOT: friday	-1.694*** (t = -9.807)	-2.089*** (t = -11.428)	-1.819*** (t = -10.510)
HOT:am630	2.426*** (t = 11.501)	3.056*** (t = 12.956)	3.224*** (t = 13.344)
HOT: am700	3.451*** (t = 14.377)	4.065*** (t = 15.268)	4.293*** (t = 16.499)
HOT: am730	3.331*** (t = 13.631)	3.981*** (t = 15.381)	3.886*** (t = 15.058)
HOT:am800	2.838*** (t = 12.061)	3.970*** (t = 16.142)	3.390*** (t = 13.740)
HOT:am830	2.463*** (t = 10.675)	2.764*** (t = 11.383)	2.458*** (t = 9.997)
HOT:am900	1.402*** (t = 6.290)	1.834*** (t = 8.204)	1.206*** (t = 5.036)
HOT:am930	-0.653*** (t = -2.947)	0.086 (t = 0.381)	-0.105 (t = -0.429)
log(hhIncomeDollars) Standard Deviation	0.423*** (t = 23.825)	0.389*** (t = 23.037)	0.378*** (t = 27.331)
HOT Share	0.502	0.526	0.555
Observations	10,000	10,000	10,000
R^2	0.370	0.328	0.398
Log Likelihood	-4,368.68	-4,650.55	-4,134.67

Figure 152 illustrates the parameter distributions that accompany this morning peak model. Notable here is the separation of the three curves. In particular, the Medium and Higher income curves do not visibly overlap with each other or with the zero value on the x-axis. The Lower segment distribution, with its mean that cannot be said to be

different from zero, straddles the zero value. The Higher income curve remains entirely in positive coefficient values, while the Medium curve remains entirely within the negative values.

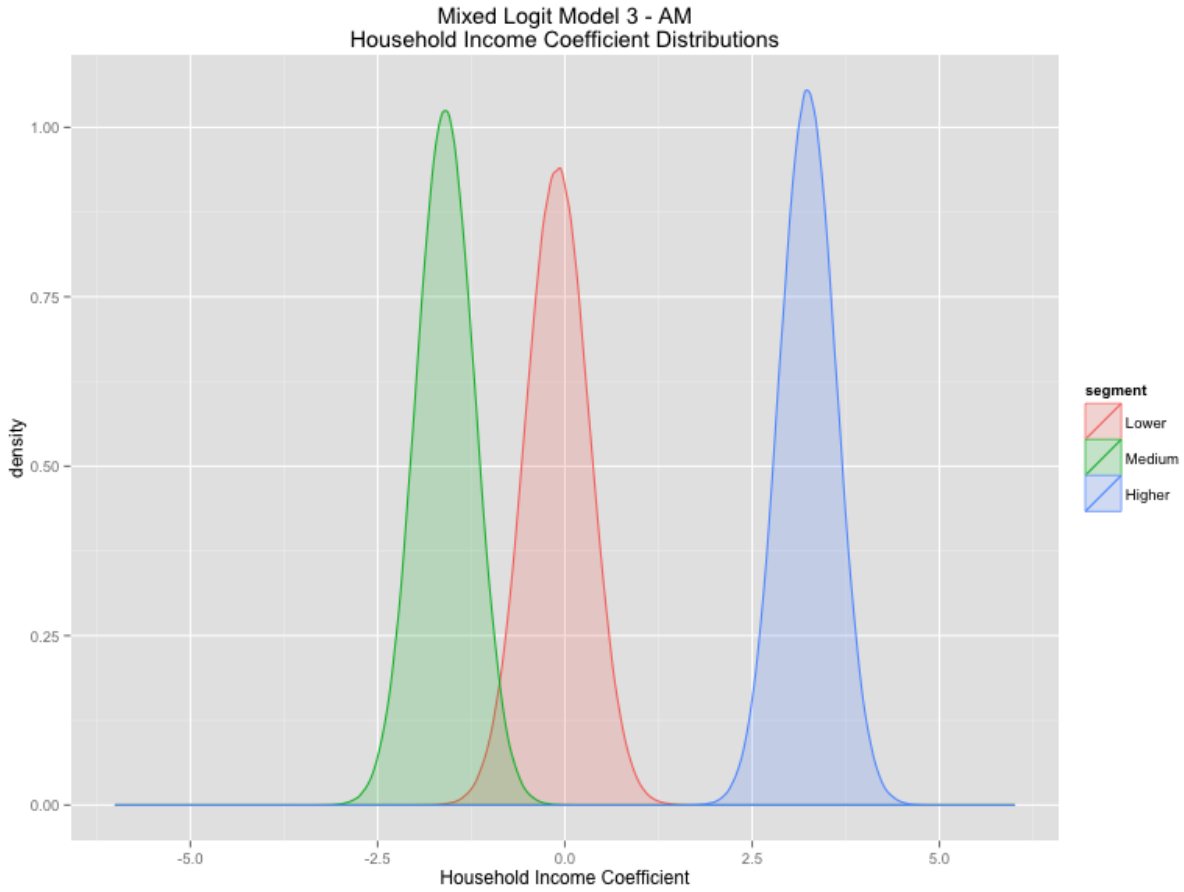


Figure 152: Normal Distributions for Household Income Parameter - AM Models

Table 110 presents the results from the afternoon peak period trip models. The relationships among these models differ from those of the morning peak: here the Lower income segment yields the only estimator that can be said to be different from zero at the 99% confidence level. The Medium income segment coefficient is also negative, while the Higher estimator is positive, but again both fail to achieve significance. As in the morning peak, the remaining factors resemble those of earlier models. For each segment, the standard deviation for the random parameter achieves significance at the 99%

confidence level, indicating that in each case the response is better modeled as a distribution.

Table 110: Mixed Logit Model 3 with 3 Income Segments – PM Peak

	PM Peak – Lower	PM Peak – Medium	PM Peak – Higher
Intercept	-0.003 (t = -0.002)	-4.63 (t = -0.885)	-8.339* (t = -1.858)
$avgSpeed^2$	-0.0003*** (t = -3.624)	-0.0004*** (t = -4.309)	-0.0001 (t = -1.429)
tollAmount	-1.080*** (t = -15.486)	-1.218*** (t = -14.738)	-0.993*** (t = -14.700)
transponderCount	0.007*** (t = 14.264)	0.011*** (t = 12.925)	0.011*** (t = 19.045)
HOT: congested40	2.689*** (t = 17.791)	3.113*** (t = 16.330)	2.455*** (t = 16.909)
HOT: hhEdu	-0.591*** (t = -7.740)	-0.510*** (t = -5.520)	-0.677*** (t = -7.596)
HOT: hhAge	-0.052 (t = -1.212)	-0.039 (t = -0.712)	-0.172*** (t = -3.017)
HOT:I(hhIncomeDollars)/hhSize)	0.00003 (t = 1.598)	0.00002 (t = 1.261)	0.00002** (t = 2.287)
HOT: log(hhIncomeDollars)	-0.585*** (t = -2.729)	-0.244 (t = -0.482)	0.243 (t = 0.581)
HOT: hhSize	0.06 (t = 0.797)	-0.022 (t = -0.267)	0.099 (t = 1.422)
HOT: segmentCount	2.529*** (t = 23.996)	2.936*** (t = 21.688)	2.673*** (t = 24.442)
HOT: february	0.567** (t = 2.233)	0.238 (t = 0.834)	-0.208 (t = -0.828)
HOT: march	0.554** (t = 2.171)	0.454 (t = 1.531)	0.186 (t = 0.748)
HOT: april	-0.173 (t = -0.687)	0.084 (t = 0.292)	-0.864*** (t = -3.220)
HOT: may	-0.603** (t = -2.292)	-0.739*** (t = -2.643)	-0.585** (t = -2.234)
HOT: june	-0.473* (t = -1.842)	-0.273 (t = -0.962)	-0.761*** (t = -2.926)
HOT: july	-0.721*** (t = -2.850)	-0.224 (t = -0.803)	-0.838*** (t = -3.027)
HOT: august	0.083 (t = 0.326)	0.206 (t = 0.727)	-0.398 (t = -1.439)
HOT: september	0.680** (t = 2.571)	1.518*** (t = 4.816)	0.344 (t = 1.261)
HOT: october	0.803*** (t = 3.060)	1.470*** (t = 4.693)	0.627** (t = 2.135)
HOT: november	0.958*** (t = 3.651)	0.954*** (t = 3.157)	0.784*** (t = 2.818)
HOT: december	0.799*** (t = 3.027)	0.954*** (t = 3.177)	0.602** (t = 2.252)
HOT: tuesday	-0.113 (t = -0.707)	-0.265 (t = -1.456)	-0.091 (t = -0.543)
HOT: wednesday	-0.149 (t = -0.916)	-0.157 (t = -0.833)	-0.109 (t = -0.660)
HOT: thursday	-0.252 (t = -1.492)	-0.271 (t = -1.382)	-0.183 (t = -1.075)
HOT: friday	-0.205 (t = -1.176)	-0.154 (t = -0.760)	-0.331* (t = -1.854)
HOT:pm1530	0.339 (t = 1.391)	0.638** (t = 2.346)	-0.051 (t = -0.215)
HOT:pm1600	0.248 (t = 1.032)	0.379 (t = 1.363)	-0.026 (t = -0.107)
HOT:pm1630	0.820*** (t = 3.496)	0.641** (t = 2.278)	0.413* (t = 1.714)
HOT:pm1700	0.897*** (t = 3.747)	1.150*** (t = 3.946)	0.696*** (t = 2.786)
HOT:pm1730	1.015*** (t = 4.202)	1.287*** (t = 4.492)	0.859*** (t = 3.677)
HOT:pm1800	1.402*** (t = 5.788)	1.598*** (t = 5.628)	1.706*** (t = 6.926)
HOT:pm1830	0.893*** (t = 3.805)	1.225*** (t = 4.532)	1.235*** (t = 5.409)
log(hhIncomeDollars) Standard Deviation	0.397*** (t = 22.813)	0.409*** (t = 19.531)	0.346*** (t = 22.585)
HOT Share	0.527	0.534	0.573
Observations	10,000	10,000	10,000
R^2	0.364	0.350	0.406
Log Likelihood	-4,400.90	-4,492.73	-4,057.75

*p<0.1; **p<0.05; ***p<0.01

Figure 153 illustrates the curves that pair with the household income parameter distributions. The three distributions all overlap each other and the zero point. The separation evident in the morning peak period chart is not present here; the response to household income across income segments is more similar in the afternoon. Another notable aspect of these curves is that all three have both positive and negative portions. Household income estimators are among those that may flip their signs from one model to the next; these results indicate that both categories of responses are appropriate for different households and conditions.

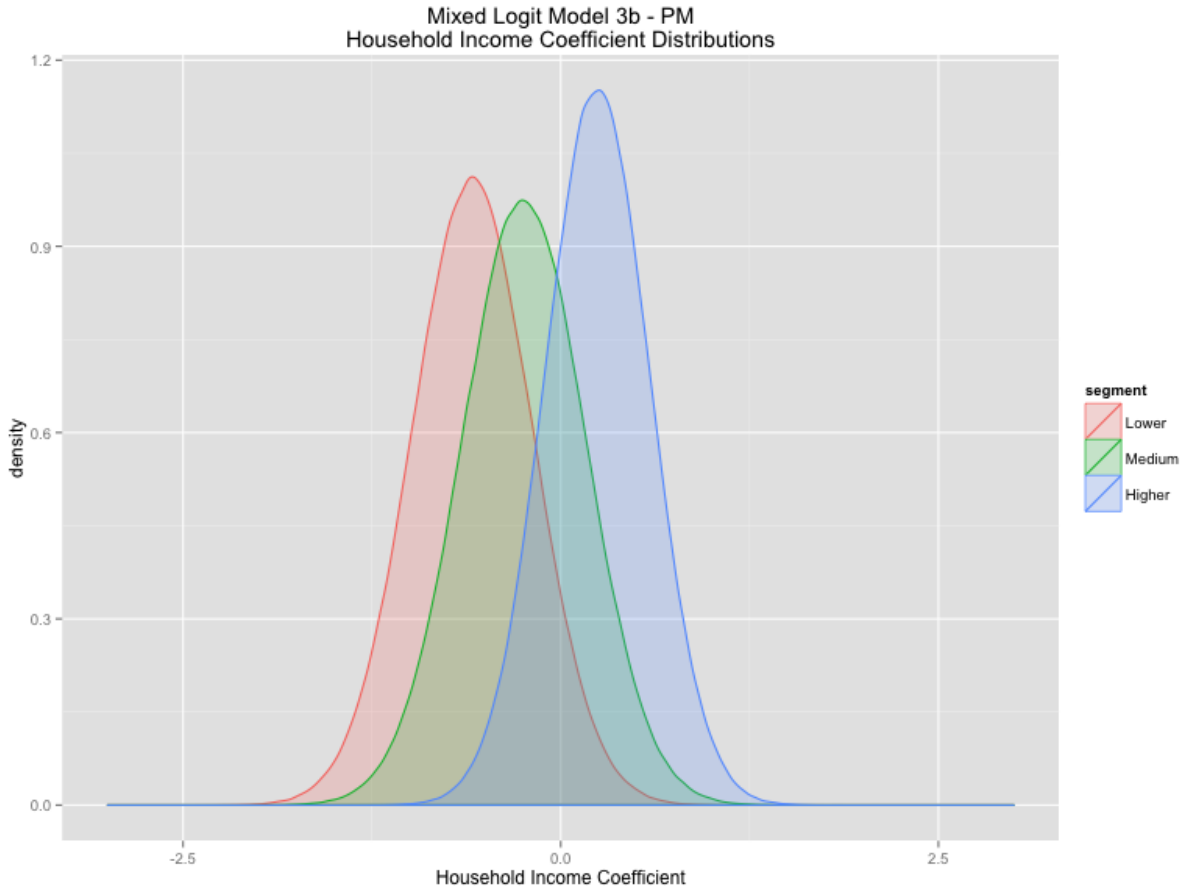


Figure 153: Normal Distributions for Household Income Parameter - PM Models

Five Segment Mixed Logit Models

As was the case with the standard binary logit models, the author found value in further sub-dividing the 'Higher' income group to better model variability in behavior among users of that group. Table 111 presents the results of the mixed logit estimation of these five income groups, with the toll amount coefficient represented by a normal distribution, for the 2013 AM peak period trips. Again, the goodness of fit improvement over the standard binary logit framework is striking. Of particular interest is the income estimator for the highest income group; it is substantially higher than those of the remaining groups. Note that the intercept term for that income group is much lower than those of the others. Table 112 similarly presents the mixed logit estimates of the PM peak period 2013 trips with the toll amount coefficient represented by a normal distribution. Again, the $200k income group yields the largest positive household income estimator, and the largest negative intercept term, but the magnitudes of these differences are smaller than their morning peak counterparts.

Figure 154 and Figure 155 present the toll amount estimator distributions for the AM and PM peak periods, respectively. In both figures, the two highest income groups ($150-200k and $200k+) have higher mean values of the toll amount coefficient than the first three groups; the $200k+ segment has the highest means of all. Like the previously estimated three-group models, the parameter distributions span positive and negative values, indicating differing impacts on Express Lane use probability due to increasing toll rates.

Table 111: Mixed Logit Model 1a with Five Income Segments - AM Peak

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	-4.475*** (t = -19.118)	-3.414*** (t = -7.482)	-7.846*** (t = -7.139)	2.241 (t = 0.543)	-31.334*** (t = -6.701)
$avgSpeed^2$	-0.0004*** (t = -54.510)	-0.001*** (t = -75.295)	-0.001*** (t = -48.866)	-0.001*** (t = -22.663)	-0.0005*** (t = -10.001)
tollAmount	-0.816*** (t = -124.283)	-0.752*** (t = -143.762)	-0.745*** (t = -89.839)	-0.332*** (t = -16.113)	-0.167*** (t = -5.188)
transponderCount	0.001*** (t = 16.756)	0.002*** (t = 53.459)	0.002*** (t = 28.070)	0.001*** (t = 5.348)	0.001*** (t = 2.814)
HOT: congested50	1.935*** (t = 84.873)	1.971*** (t = 102.604)	2.124*** (t = 68.912)	2.557*** (t = 33.351)	2.250*** (t = 20.077)
HOT: hhEdu	0.030*** (t = 3.414)	-0.034*** (t = -4.300)	-0.165*** (t = -11.931)	-0.014 (t = -0.385)	0.179*** (t = 3.893)
HOT: hhAge	-0.028*** (t = -5.212)	0.028*** (t = 5.813)	-0.092*** (t = -9.960)	-0.047* (t = -1.942)	-0.187*** (t = -4.666)
HOT:I(hhIncomeDollars)/hhSize)	-0.00000* (t = -1.780)	0.00000*** (t = 3.104)	0 (t = 0.564)	0.00001*** (t = 2.932)	-0.00002*** (t = -5.018)
HOT: log(hhIncomeDollars)	0.118*** (t = 4.484)	0.025 (t = 0.572)	0.507*** (t = 5.207)	-0.532 (t = -1.510)	2.528*** (t = 6.537)
HOT: hhSize	-0.041*** (t = -4.593)	-0.016** (t = -2.187)	-0.028*** (t = -2.665)	0.096*** (t = 3.113)	-0.290*** (t = -6.846)
HOT: segmentCount	1.030*** (t = 142.041)	1.037*** (t = 176.218)	1.051*** (t = 108.828)	1.031*** (t = 43.094)	0.888*** (t = 24.683)
HOT: february	0.456*** (t = 14.371)	0.421*** (t = 16.209)	0.492*** (t = 12.055)	0.308*** (t = 3.126)	0.269* (t = 1.846)
HOT: march	0.465*** (t = 14.864)	0.496*** (t = 19.154)	0.458*** (t = 11.463)	0.404*** (t = 4.123)	0.586*** (t = 4.097)
HOT: april	0.675*** (t = 21.324)	0.661*** (t = 24.948)	0.684*** (t = 16.851)	0.471*** (t = 4.657)	0.684*** (t = 4.819)
HOT: may	0.526*** (t = 17.499)	0.507*** (t = 20.235)	0.662*** (t = 16.869)	0.578*** (t = 6.058)	0.534*** (t = 4.011)
HOT: june	0.535*** (t = 16.642)	0.525*** (t = 19.435)	0.603*** (t = 14.638)	0.720*** (t = 6.889)	0.441*** (t = 3.057)
HOT: july	0.309*** (t = 10.249)	0.333*** (t = 13.354)	0.395*** (t = 9.933)	0.430*** (t = 4.549)	0.253* (t = 1.868)
HOT: august	0.627*** (t = 19.864)	0.600*** (t = 23.144)	0.718*** (t = 17.689)	0.459*** (t = 4.744)	0.261* (t = 1.844)
HOT: september	0.704*** (t = 22.223)	0.646*** (t = 24.841)	0.718*** (t = 17.147)	0.728*** (t = 6.908)	0.442*** (t = 2.818)
HOT: october	0.772*** (t = 24.511)	0.711*** (t = 27.397)	0.768*** (t = 18.395)	0.693*** (t = 6.748)	0.565*** (t = 3.603)
HOT: november	0.459*** (t = 14.981)	0.485*** (t = 19.155)	0.495*** (t = 12.157)	0.491*** (t = 4.976)	0.417*** (t = 2.903)
HOT: december	0.050* (t = 1.684)	0.069*** (t = 2.807)	0.150*** (t = 3.849)	0.115 (t = 1.240)	-0.048 (t = -0.351)
HOT: tuesday	0.389*** (t = 18.367)	0.330*** (t = 18.753)	0.302*** (t = 10.859)	0.359*** (t = 5.314)	0.476*** (t = 4.770)
HOT: wednesday	0.480*** (t = 22.338)	0.369*** (t = 20.558)	0.379*** (t = 13.424)	0.461*** (t = 6.624)	0.562*** (t = 5.414)
HOT: thursday	0.411*** (t = 19.509)	0.375*** (t = 21.351)	0.351*** (t = 12.739)	0.478*** (t = 6.840)	0.548*** (t = 5.390)
HOT: friday	-0.798*** (t = -39.095)	-0.808*** (t = -47.771)	-0.917*** (t = -34.173)	-0.813*** (t = -12.234)	-0.349*** (t = -3.710)
HOT:am630	1.762*** (t = 61.354)	1.684*** (t = 71.363)	1.818*** (t = 46.075)	2.016*** (t = 19.711)	0.913*** (t = 6.748)
HOT: am700	1.932*** (t = 65.404)	1.861*** (t = 76.617)	2.124*** (t = 54.196)	2.297*** (t = 23.070)	1.432*** (t = 9.721)
HOT: am730	1.908*** (t = 64.257)	1.652*** (t = 68.969)	1.872*** (t = 47.462)	1.930*** (t = 19.790)	1.150*** (t = 8.069)
HOT:am800	1.460*** (t = 49.912)	1.258*** (t = 53.100)	1.316*** (t = 33.391)	1.398*** (t = 14.283)	0.838*** (t = 6.170)
HOT:am830	1.017*** (t = 35.443)	0.835*** (t = 35.704)	0.827*** (t = 21.384)	1.073*** (t = 11.607)	0.203 (t = 1.549)
HOT:am900	0.375*** (t = 13.585)	0.138*** (t = 6.286)	0.208*** (t = 5.751)	0.388*** (t = 4.356)	-0.423*** (t = -3.380)

Table 111 Continued

HOT:am930	-0.482*** (t = -18.311)	-0.590*** (t = -28.349)	-0.546*** (t = -15.743)	-0.420*** (t = -4.937)	-0.888*** (t = -7.796)
HOT Share	0.507	0.520	0.538	0.617	0.673
Observations	342,209	533,623	224,862	39,610	19,670
R^2	0.599	0.619	0.63	0.626	0.616
Log Likelihood	-95,048.05	-140,855.00	-57,495.02	-9,864.56	-4,770.99

Table 112: Mixed Logit Model 1a with Five Income Segments - PM Peak

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	-4.171*** (t = -16.276)	-2.753*** (t = -5.849)	3.902*** (t = 3.488)	-2.067 (t = -0.522)	-13.938*** (t = -3.208)
$avgSpeed^2$	-0.0002*** (t = -17.749)	-0.0002*** (t = -17.281)	-0.0002*** (t = -11.321)	-0.0001*** (t = -3.170)	-0.0001 (t = -1.360)
tollAmount	-0.359*** (t = -43.639)	-0.289*** (t = -44.265)	-0.295*** (t = -28.974)	0.029 (t = 1.208)	0.516*** (t = 11.468)
transponderCount	0.007*** (t = 102.907)	0.008*** (t = 155.630)	0.009*** (t = 104.981)	0.008*** (t = 43.813)	0.007*** (t = 25.530)
HOT: congested40	1.709*** (t = 98.157)	1.687*** (t = 116.033)	1.679*** (t = 73.861)	1.543*** (t = 30.125)	1.641*** (t = 20.569)
HOT: hhEdu	0.018* (t = 1.922)	-0.031*** (t = -3.869)	0.021 (t = 1.544)	-0.171*** (t = -5.052)	-0.102** (t = -2.122)
HOT: hhAge	0.003 (t = 0.517)	-0.001 (t = -0.303)	0.011 (t = 1.227)	-0.116*** (t = -5.305)	-0.019 (t = -0.476)
HOT:I(hhIncomeDollars)/hhSize)	0.00001*** (t = 2.931)	0.00001*** (t = 5.600)	0.00000* (t = 1.815)	-0.00001*** (t = -4.265)	-0.00001** (t = -2.234)
HOT: log(hhIncomeDollars)	-0.045 (t = -1.576)	-0.190*** (t = -4.165)	-0.789*** (t = -7.976)	-0.043 (t = -0.127)	0.803** (t = 2.265)
HOT: hhSize	0.014 (t = 1.409)	0.034*** (t = 4.650)	0.043*** (t = 4.035)	-0.086*** (t = -3.397)	-0.036 (t = -0.932)
HOT: segmentCount	1.511*** (t = 183.759)	1.587*** (t = 234.261)	1.677*** (t = 154.242)	1.610*** (t = 65.594)	1.597*** (t = 43.566)
HOT: february	-0.100*** (t = -3.182)	-0.086*** (t = -3.278)	-0.155*** (t = -3.786)	-0.095 (t = -1.060)	-0.185 (t = -1.361)
HOT: march	-0.088*** (t = -2.792)	-0.087*** (t = -3.315)	-0.146*** (t = -3.606)	-0.006 (t = -0.064)	-0.147 (t = -1.065)
HOT: april	-0.276*** (t = -8.700)	-0.284*** (t = -10.777)	-0.369*** (t = -8.917)	-0.374*** (t = -4.059)	-0.535*** (t = -3.884)
HOT: may	-0.472*** (t = -15.123)	-0.419*** (t = -16.190)	-0.507*** (t = -12.621)	-0.325*** (t = -3.651)	-0.521*** (t = -3.892)
HOT: june	-0.471*** (t = -14.847)	-0.436*** (t = -16.323)	-0.535*** (t = -12.786)	-0.303*** (t = -3.203)	-0.736*** (t = -5.068)
HOT: july	-0.467*** (t = -15.092)	-0.419*** (t = -16.157)	-0.551*** (t = -13.571)	-0.312*** (t = -3.372)	-0.671*** (t = -4.914)
HOT: august	-0.213*** (t = -6.594)	-0.125*** (t = -4.648)	-0.249*** (t = -5.960)	-0.066 (t = -0.709)	-0.638*** (t = -4.303)
HOT: september	0.236*** (t = 7.098)	0.426*** (t = 15.557)	0.260*** (t = 6.090)	0.350*** (t = 3.730)	-0.354** (t = -2.447)
HOT: october	0.310*** (t = 9.294)	0.532*** (t = 19.224)	0.354*** (t = 8.129)	0.282*** (t = 2.896)	-0.173 (t = -1.214)
HOT: november	0.176*** (t = 5.343)	0.386*** (t = 14.092)	0.313*** (t = 7.409)	0.222** (t = 2.316)	-0.412*** (t = -2.846)
HOT: december	0.091*** (t = 2.810)	0.266*** (t = 9.757)	0.284*** (t = 6.670)	0.151 (t = 1.538)	-0.269* (t = -1.898)
HOT: tuesday	-0.012 (t = -0.578)	0.011 (t = 0.613)	-0.055** (t = -2.032)	-0.007 (t = -0.113)	0.03 (t = 0.323)
HOT: wednesday	0.02 (t = 0.941)	0.016 (t = 0.907)	-0.013 (t = -0.469)	0.032 (t = 0.503)	-0.023 (t = -0.245)
HOT: thursday	-0.131*** (t = -6.023)	-0.145*** (t = -7.953)	-0.189*** (t = -6.704)	-0.256*** (t = -3.982)	-0.198** (t = -2.086)
HOT: friday	-0.072*** (t = -3.213)	-0.144*** (t = -7.710)	-0.118*** (t = -4.029)	-0.156** (t = -2.319)	0.177* (t = 1.777)
HOT: pm1530	-0.088*** (t = -3.140)	-0.120*** (t = -5.153)	-0.178*** (t = -5.051)	-0.376*** (t = -4.890)	0.061 (t = 0.557)
HOT: pm1600	-0.160*** (t = -5.519)	-0.195*** (t = -8.237)	-0.120*** (t = -3.311)	-0.209*** (t = -2.622)	-0.017 (t = -0.145)
HOT: pm1630	0.036 (t = 1.237)	0.038 (t = 1.585)	0.188*** (t = 5.135)	0.052 (t = 0.659)	0.211* (t = 1.686)
HOT: pm1700	0.214*** (t = 7.193)	0.294*** (t = 12.057)	0.408*** (t = 10.809)	0.216** (t = 2.572)	0.215* (t = 1.688)
HOT: pm1730	0.400*** (t = 13.457)	0.509*** (t = 20.793)	0.635*** (t = 16.893)	0.577*** (t = 6.902)	0.760*** (t = 6.316)
HOT: pm1800	0.646*** (t = 22.531)	0.652*** (t = 27.309)	0.793*** (t = 21.191)	0.633*** (t = 7.901)	0.709*** (t = 6.157)

Table 112 Continued

HOT: pm1830	0.521*** (t = 18.764)	0.550*** (t = 23.691)	0.633*** (t = 17.776)	0.368*** (t = 4.637)	0.743*** (t = 6.704)
HOT Share	348,894	544,660	235,228	43,388	21,138
Observations	0.601	0.625	0.632	0.599	0.609
R^2	-96,320.97	-140,890.70	-59,356.57	-11,825.64	-5,569.71
Log Likelihood	348,894	544,660	235,228	43,388	21,138

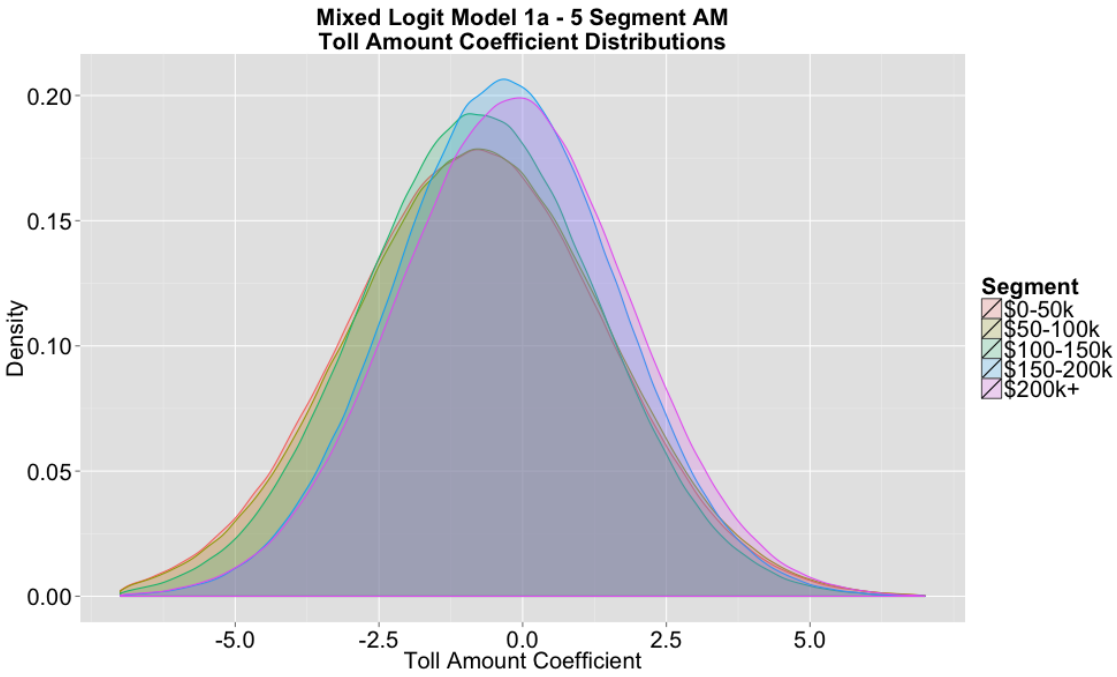


Figure 154: Normal Distributions for Toll Amount Parameter - 5 Segment AM Models

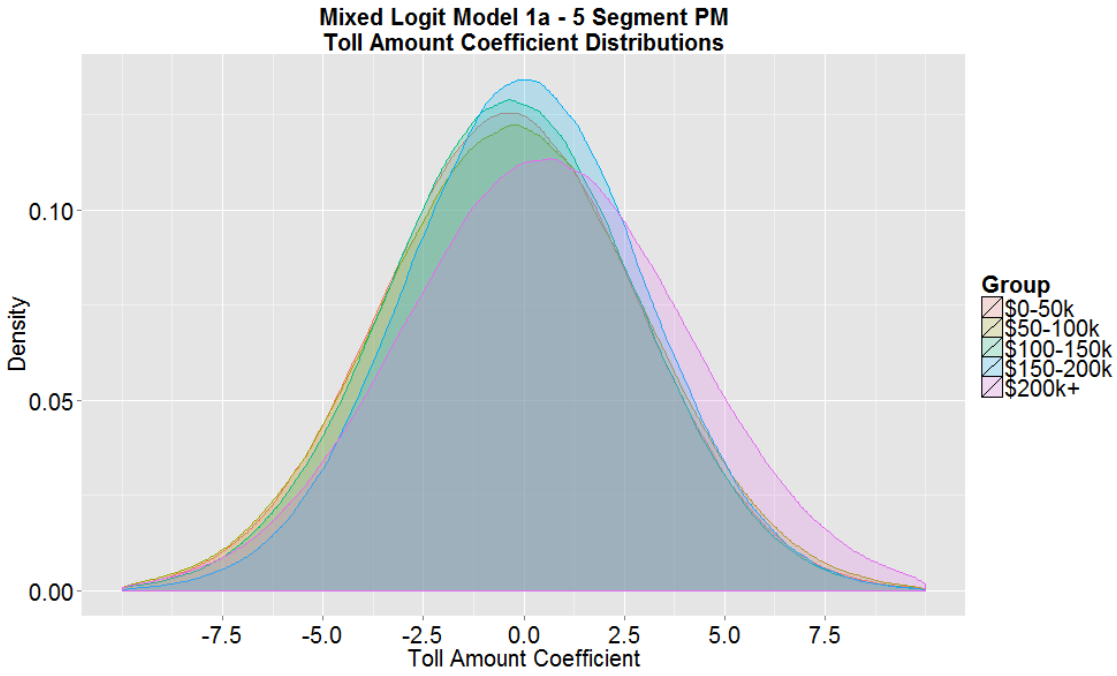


Figure 155: Normal Distributions for Toll Amount Parameter - 5 Segment PM Models

Demand Elasticity Results

As discussed in the Literature Review chapter of this dissertation, elasticity is a measure of the relative impact on one measure based on the change in another measure. The measure is reported as a value that represents a percentage change. If y has an elasticity with respect to x of 1.1, for example, that means that a 1% change in x results in a 1.1% change in y. In this analysis, the dependent variable is the probability of choosing the Express Lanes on a given trip, while the independent variables are numerous and presented below in the various charts. This section uses Model 14b, with both the three income segment and five income segment methods, as the basis for its elasticity analysis. The values represented in the charts represent the average of all of the disaggregate elasticity values derived from the logit models.

Figure 156 presents the results from the morning peak period trips, estimated for three income segments. All three segments exhibit nearly unitary elasticity with respect to toll amount. A unitary elasticity value (a value of one, or in this case, negative one) indicates that a 1% increase in the toll level results in a 1% decrease in a user's probability of choosing the Express Lanes for a trip. The only elasticity value that exceeds one is that of segment count: it is consistently the highest elasticity result across all income segments. Household education yields negative elasticity values of similar magnitudes across all three segments. Only the Higher income segment exhibits positive sensitivity to household income; this result is explored later in the five-segment model elasticity results.

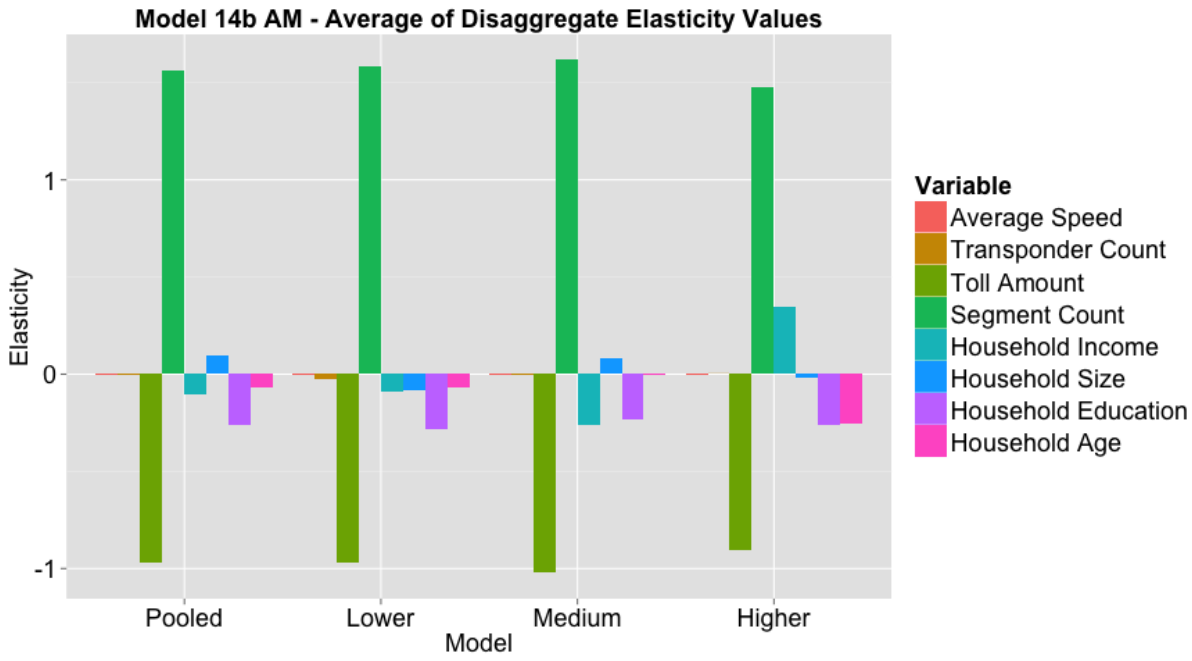


Figure 156: Elasticity Values - Three Segments - AM

The averages of the disaggregate elasticity values for the afternoon peak models are shown below in Figure 157. Again, the segmentCount variable yields the highest elasticity value across all segments. It should be noted that when the distance variable was included in the earlier models, its average elasticity value was also consistently the highest across both time periods and all three income segments. Toll amount sensitivities are much lower in the afternoon peak: while they were at or near negative one in the morning, here none of the models report elasticities that exceed -0.5. Household education elasticity levels are also consistently negative, and in fact exceed the toll amount levels for all segments. Unlike the morning trips, the household income factor hast little impact on the afternoon trip decision making process. Other variables yield elasticity values that are very close to zero.

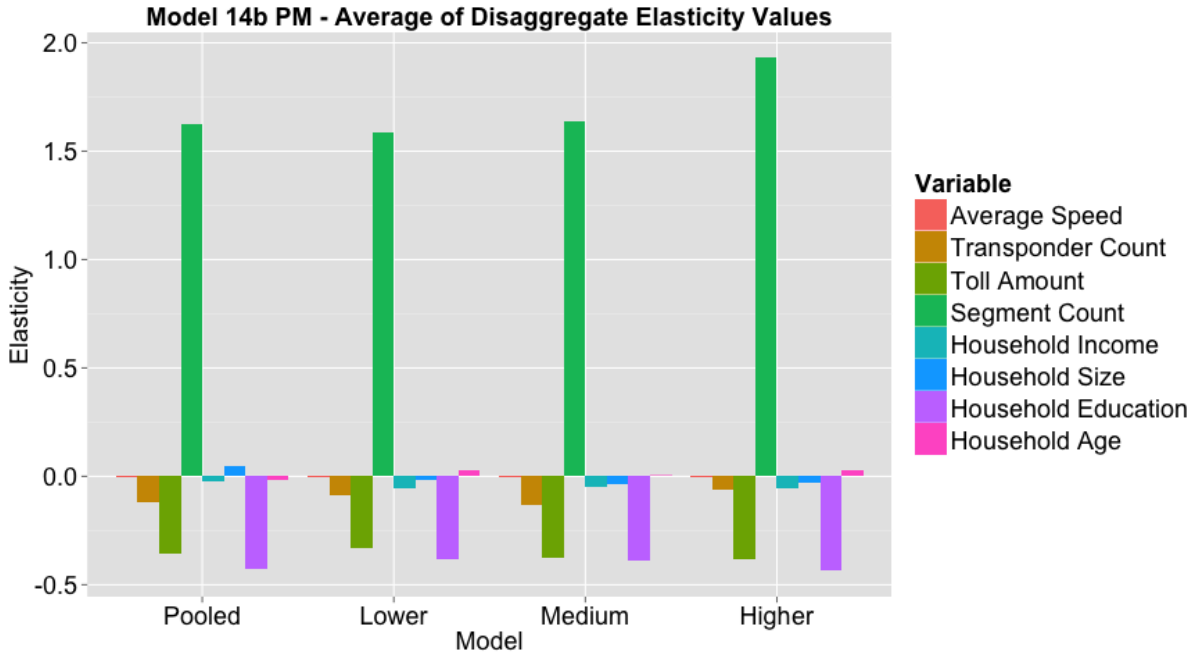


Figure 157: Elasticity Values - Three Segments - PM

The next pair of figures illustrate the demand elasticity results generated by the five-segment estimates of Model 14b. Figure 158 presents the morning peak period results. The two additional segments within the earlier Higher income segment demonstrate substantial differences relative to the three original segments and to the Higher income segment specifically. The two factors that dominate the elasticity results of the previous morning peak models, segment count and toll amount, diminish in magnitude as segment income increases beyond the $100-150k category. Within the highest income segment, that of households making $200k+, household age is the factor that households are the most sensitive to. The household income, size, and education factors all yield negative elasticity values in the highest segment, though their magnitudes are well below one. The $150-200k segment is unique in that it has the greatest

sensitivity to household education levels, and also the only positive sensitivity to household income within the afternoon peak models.

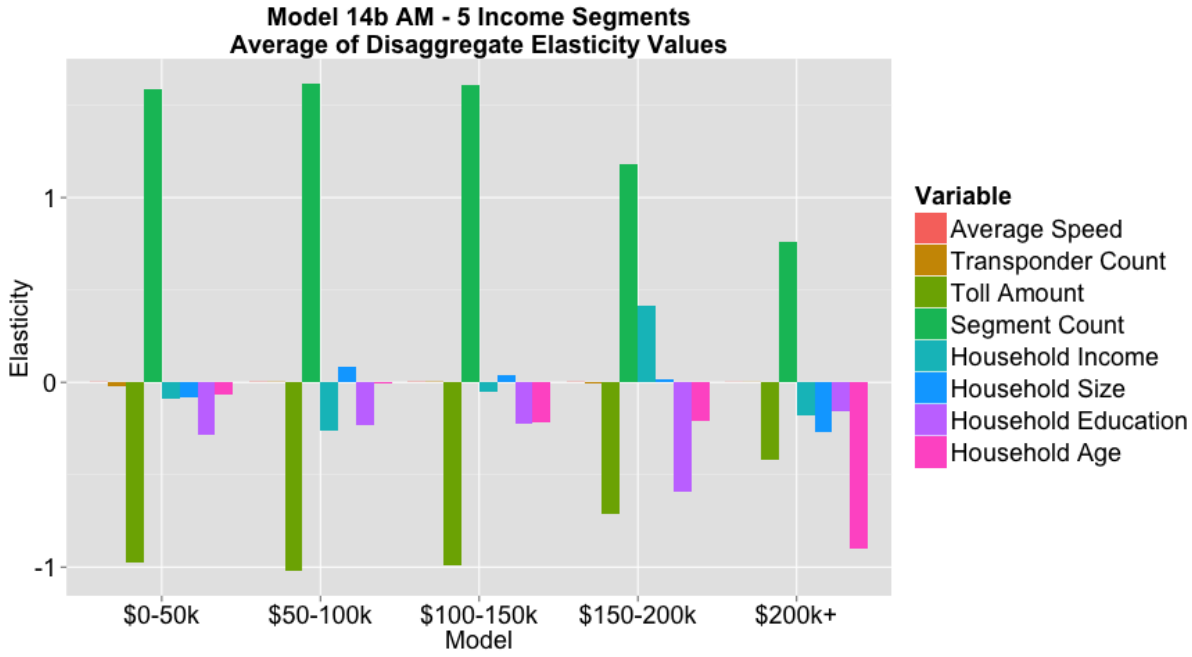


Figure 158: Elasticity Values - Five Segments - AM

Figure 159 illustrates the five-segment elasticity results for the afternoon peak period. The segment of users with over $200,000 in annual household income has the largest elasticity value with regards to household income and the lowest sensitivity to trip segment counts. In the three-segment analysis, the income effect in the Higher segment was very close to zero; the aggregation of the households within that segment disguised the behavioral variation within it. The highest income segment here is also notable for having the lowest elasticity with respect to toll amount, though the toll elasticities for all of the segments are well below unitary.

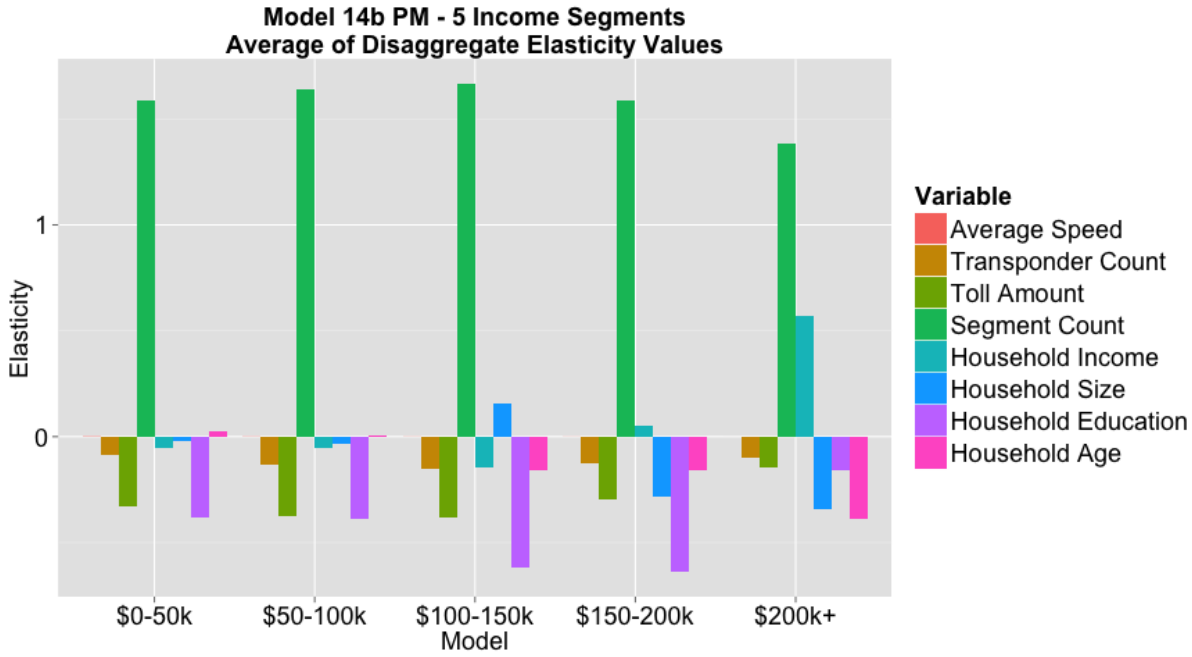


Figure 159: Elasticity Values - Five Segments – PM

Finally, Figure 160 illustrates the averages of the disaggregate elasticity values for the morning peak period mixed logit models with five income groups. Toll amount elasticities are similar for the first three income groups; the lowest income group is slightly more sensitive to toll rates in the mixed logit model versus the standard logit model. The two highest income groups see less of an impact on Express Lane use probability as toll rates increase: the toll amount elasticities for those two groups are closer to zero than in the standard logit framework. Segment count elasticity is higher across all five income groups with the mixed logit framework. The highest income group, $200k+, also has a much higher elasticity with respect to household income than in the previous models.

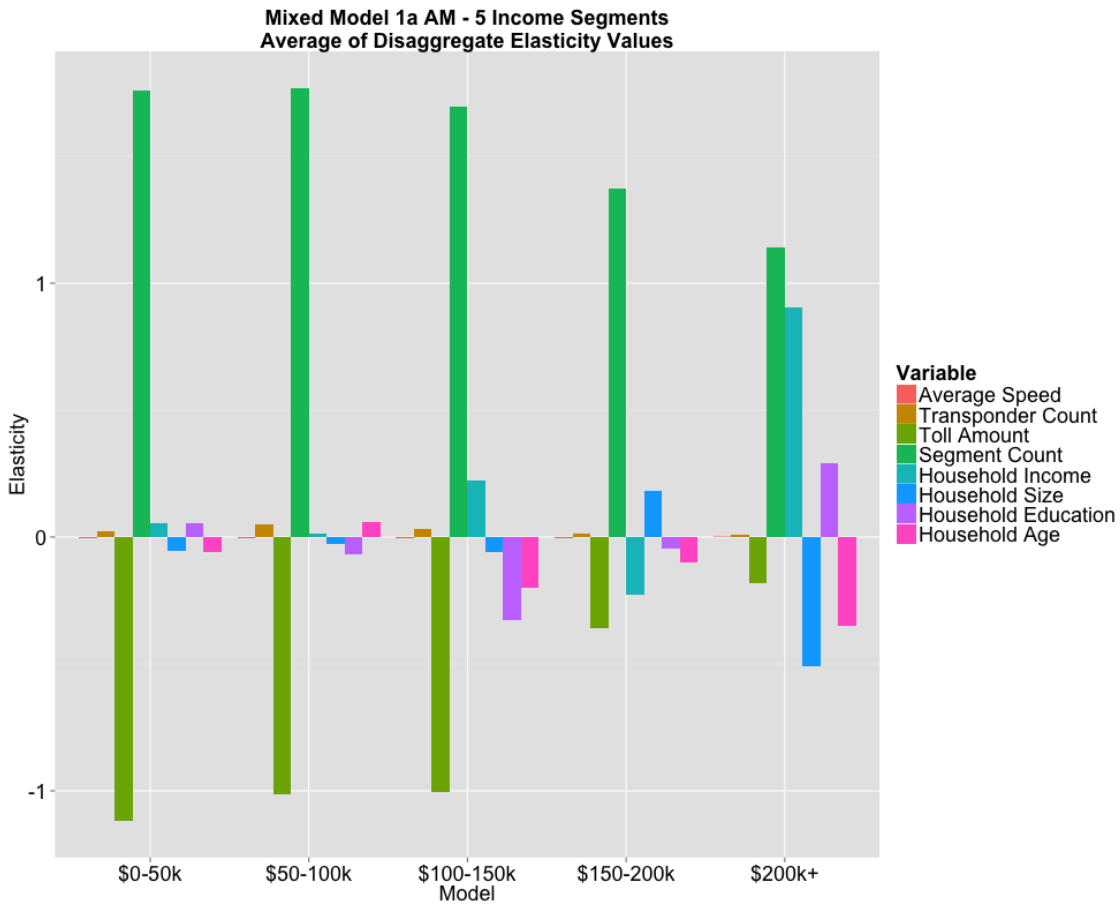


Figure 160: Mixed Logit Elasticity Values - 5 Segments - AM

Price Elasticity For Different Toll Amounts

One shortcoming of the averages of disaggregate elasticity values presented in the preceding figures is their inability to demonstrate the sensitivities of a variable to a range of values. Demand elasticity at a price of $10, for example, may not be the same as the elasticity when the price is $1. The next series of figures investigates toll rate elasticity across the range of potential toll amounts for users in the income segments defined above. In these charts, the values of the other factors in the model were all set to the mean values in the data set.

Figure 161 shows the results for the three-segment Model 14b in the morning peak period. Each segment exhibits different patterns of toll elasticity. For much of the

range of potential toll values, until nearly $4, the Medium income segment exhibits elasticity values very close to zero. At that point the users begin displaying more sensitivity to toll amounts, though they remain the least elastic of the three segments. The Lower segment users are also inelastic until roughly the $1 mark; after that their curve is steeper than that of the Medium segment. The Higher income users exhibit the highest and most consistently increasing levels of price elasticity; this may be an artifact of the remaining factors held constantly at their means.

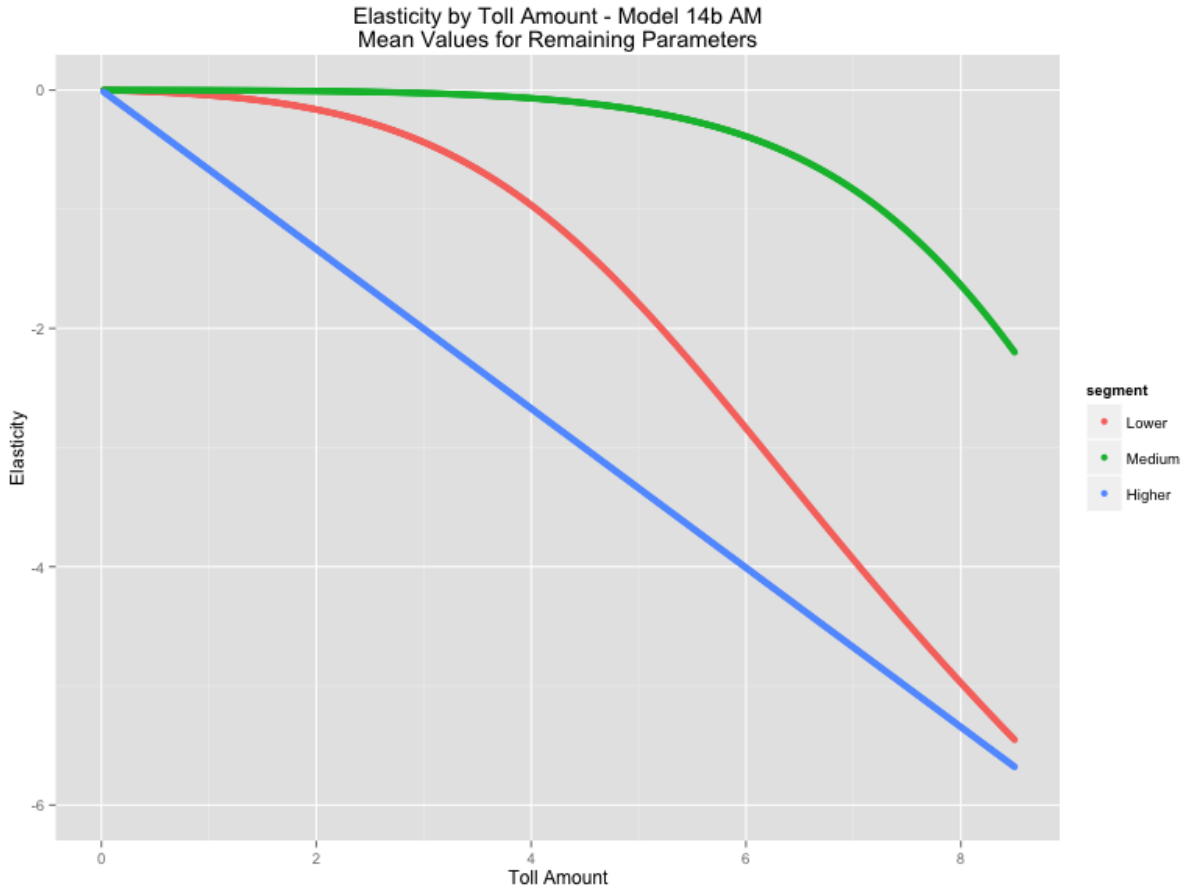


Figure 161: Elasticity by Toll Amount - Three Segments - AM

A similar pattern can be observed in Figure 162, which shows the elasticity values across the toll range of the afternoon peak period trips. The pattern here is similar: the Higher income segment exhibits a nearly linear rate of change in price demand elasticity, while the Lower and Medium segments are slower to increase their sensitivity magnitudes. In all three cases, the afternoon elasticity values are lower in magnitude than the morning values: no segment exceeds an elasticity value of -4 in the afternoon, while two segments have final elasticity values exceeding -5 in the morning peak.

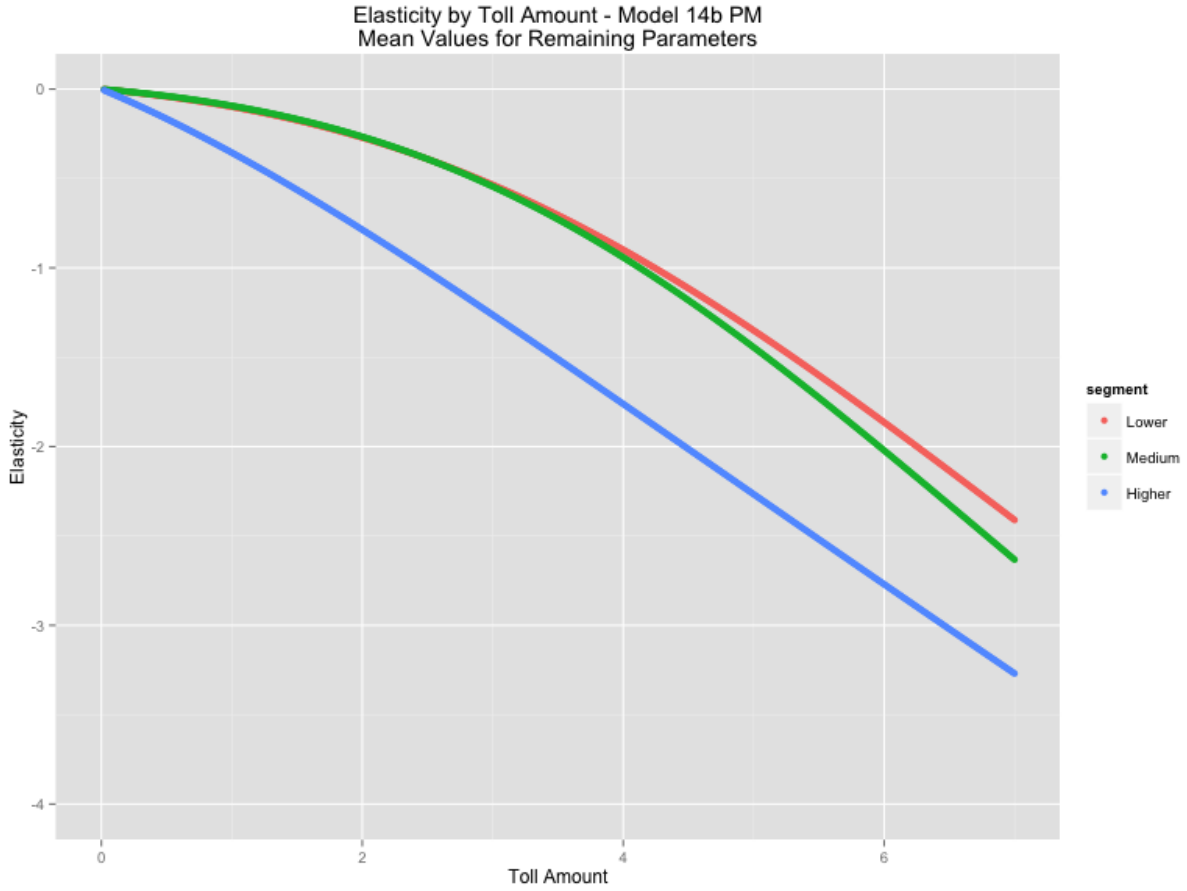


Figure 162: Elasticity by Toll Amount - Three Segments - PM

The previous modeling investigations revealed the benefits of further sub-dividing the Higher income segment into smaller categories. Figure 163 shows the elasticity ranges of those sub-segments for the 2013 morning peak period. Whereas in Figure 161 the Higher income segment had a constant rate of change with regards to its toll demand elasticity, Figure 163 reveals the variety of responses within that category. The highest income segment, representing households with over $200,000 in annual income, is nearly perfectly inelastic across the entire range of toll amounts. The $150-200k segment most closely resembles the linear curve seen in Figure 161. The $100-150k segment exhibits a slower rate of elasticity change at lower toll amounts, relative to the $150-200k segment, and then yields a steeper curve for higher prices.

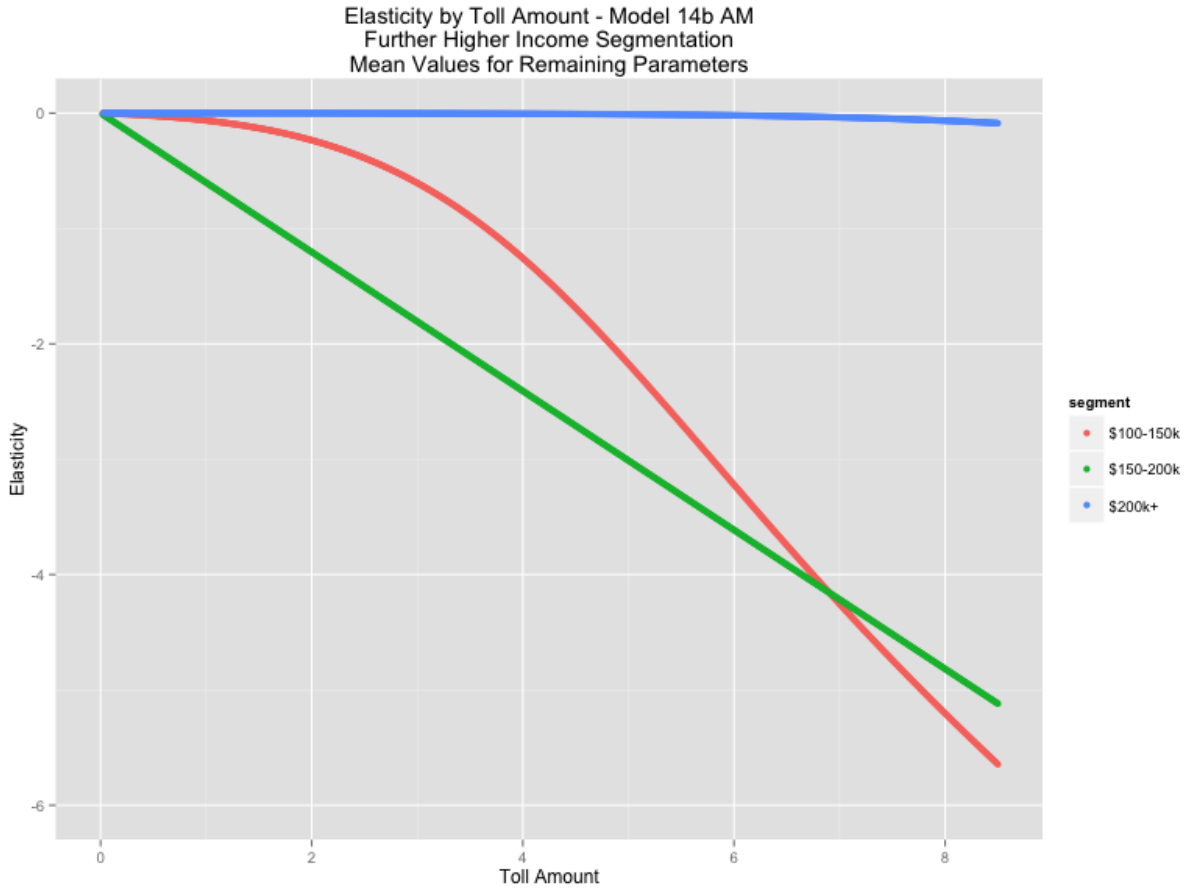


Figure 163: Elasticity by Toll Amount - Higher Income Segments – AM

Finally, Figure 164 shows the afternoon peak elasticity curves for the sub-segments of the Higher income group. The $200k+ segment behaves very differently in the afternoon, in that those users now exhibit a constant increase in elasticity, the magnitudes of which exceed those of the $100-150k segment. The $150-200k segment is most sensitive to toll amounts higher than $2.50. The lowest segment, $100-150k, has the lowest price elasticity response: its non-linear curve only exceeds -1 after the price exceeds $6.

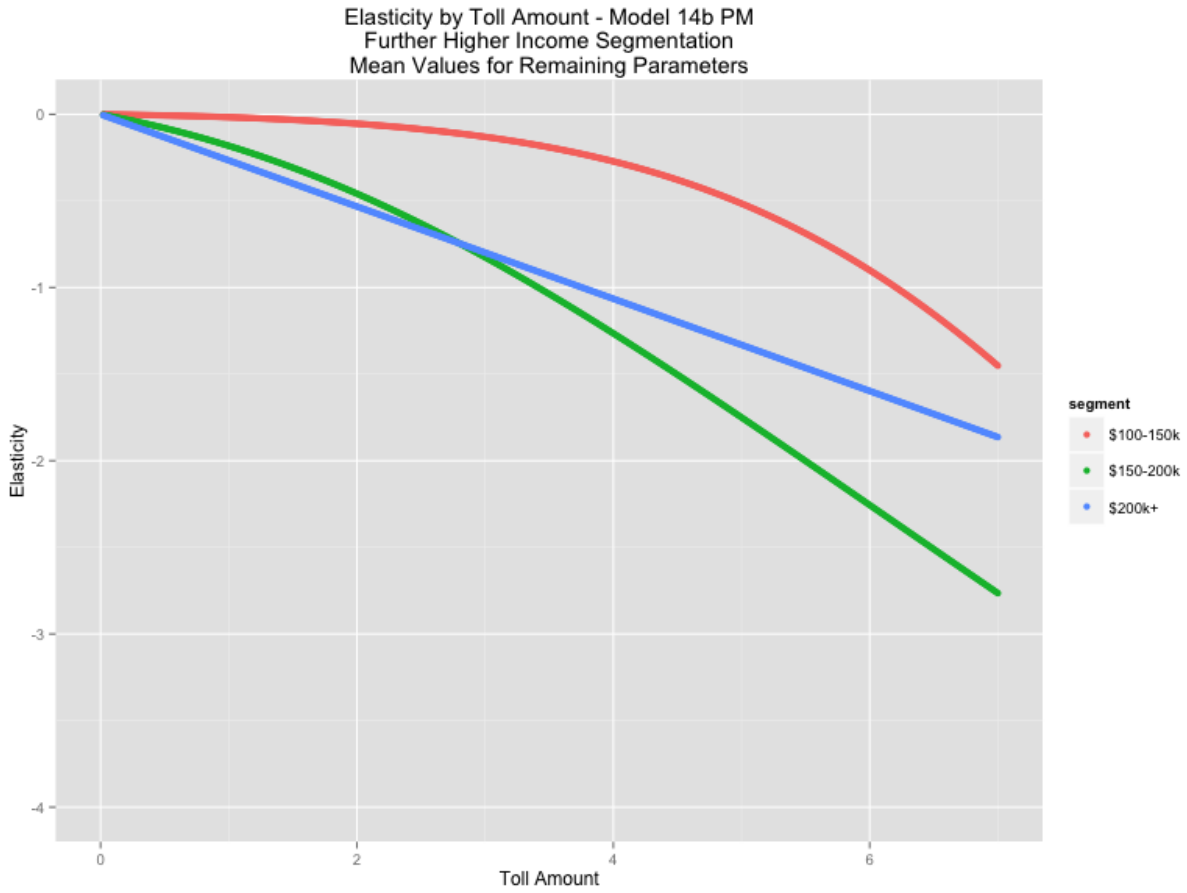


Figure 164: Elasticity by Toll Amount - Higher Income Segments - PM

Chapter Overview

This chapter sought to improve upon the initial modeling work presented in Chapter 8 by expanding the data set, adding new variables and interaction terms, and investigating different methods to address the shortcomings of the earlier analyses. These shortcomings included serial correlation caused by multiple observations of the same user, the lack of partial-corridor trips, and the decrease in model performance that resulted from the aggregation of morning and afternoon trips as well as high income households. The chapter also examined the random parameter distributions and elasticity values that were generated from the various models.

The model building and selection process involved over twenty different models with varying factors and interaction terms. The final models were selected based on coefficient significance, behavioral characteristics, goodness of fit, and the Akaike information criterion measure of model quality. The investigation revealed the benefits of separately modeling morning and afternoon peak period trips, and also of further segmenting households by income to illustrate the variety of behavior within the higher income households. This segmentation indicated that the three-segment strategy disguised substantial behavioral differences among the highest income households on the I-85 corridor. The determinants of lane choice decision-making in the morning peak had notable differences from the determinants of the afternoon peak, particularly with regards to toll rate sensitivity and the impact of the total corridor segments traversed. Afternoon peak models had better goodness of fit metrics overall, though the pseudo-R^2 measures for both time frames were under 0.400 in all but one of the cases.

The mixed logit framework improved the modeling results by addressing the serial correlation that resulted from the panel data used in the analysis. Estimating the toll amount and household income coefficients as random rather than fixed parameters provided evidence for the varying nature of the impacts of these factors on lane choice decisions by households. The toll amount coefficients, for example, were more appropriately modeled as normal distributions that encapsulate both positive and negative values to reflect both the 'signaling' and demand-reducing effects of toll rates.

Further segmenting the study households showed that lane choice determinants varied more within the 'Higher' income segment than across the original three-segment structure. In particular, the five-segment models illustrated lower elasticities with regard to corridor segment counts and toll levels for the highest-income households in the sample, as well as higher household income level elasticities for afternoon trips by that same cohort.

The models estimated in this chapter also allowed for the measure of price elasticity of demand across the spectrum of toll rates charged on the corridor. The results of this analysis suggest a lack of consistent elasticity patterns using these models, especially with regards to the morning and afternoon curves for the same sets of users. Serial correlation likely biases these results as they do not originate from the mixed logit models. Further investigation may potentially reveal different sensitivities with different parameter values; for example, researchers may use the median rather than mean values for the remaining factors, or look at different segment counts in isolation.

Despite the enhancements made to the preliminary models in this chapter, further model development remains both possible and desirable. While the set of variables used

in this analysis greatly expanded upon those used in the preliminary modeling efforts, further complications arose in the form of collinearity among the operational factors. The transponder count factor was included to provide some measure of the overall demand, though a more comprehensive measure that was not restricted to Peach Pass-holding vehicles would be preferred. Estimating the five-segment models with the mixed logit framework, while time consuming, would likely provide more insights. Finally, addressing the issues of sample bias and match rates between the SRTA and Epsilon data, outlined earlier in this dissertation, could improve models and provide a more comprehensive overview of the users and non-users of the Express Lanes.

CHAPTER 13

CONCLUSION

The I-85 Express Lanes represent the first step in a planned $16 billion investment in value-priced facilities in the Atlanta region. This analysis of user response with regards to value of time, lane choice decision determinants, and demand elasticity has important implications for demand management and equity analysis. The research conducted here had a number of unique characteristics. The data set is a combination of two sources that are not typically seen in pricing research: disaggregated, automated Express Lane use and non-use data and privately sourced household level socioeconomic data. The methods included both familiar studies and innovative uses of the data. The resulting analysis had four different objectives: to measure value of time and price demand elasticity by examining the revealed behavior of toll lane users, to improve understanding of individual-level lane choice decisions by examining the determinants of lane use, to use new and unique data sources to improve modeling outcomes, and to compare the effects of trip characteristics among different population segments. This was accomplished by examining users' value of travel time savings and price elasticity of demand, and assessing the potential determinants of Express Lane decision making for different population segments. These data sources, methods, and objectives served to answer the question of how consumers respond differently to road pricing based on operational and demographic differences. The research involved in this dissertation fell into three broad categories: investigating the data loss and sample bias that arose from the data processing methods, examining the value of travel time savings exhibited by users of the Express

Lanes, and modeling lane choice behavior with a combination of demographic characteristics and corridor conditions.

Research Findings

Value of Travel Time Savings

This dissertation used the revealed preference data of I-85 Express Lane users to investigate the monetary value users ascribed to their time on the corridor, by examining the toll amounts they paid and the resulting time that they saved. The analyses examined the resulting value of travel time savings distributions across income segments and among trips of different lengths. The differences in these distributions among lower, medium, and higher income households were marginal at best. Differences among the mean, median, and other quartile values were on the order of cents rather than dollars. The results did not indicate that higher income households had the highest value of travel time savings results, as may have been expected. The ranking of VTTS values by income segment was not consistent across time frames or directions. The trip length investigation revealed more distinct differences between users who traverse the entire duration of the corridor and those that take partial trips; in that case, the southbound and northbound differences were also more pronounced. An important consideration in interpreting these results is that they represent the Express Lane users only; that is, only users who chose to make paid trips in the HOT lanes. Non-users, and general purpose lane trips by HOT users, were excluded from this analysis.

HOT Lane Choice Modeling

The modeling work performed here provided a number of insights into toll lane use and the determinants of lane choice decisions. The discrete choice analysis was performed in two phases, the first of which will be published in the Transportation Research Record. This preliminary analysis was extended with additional variables, observations, and methods, the results of which were included in this dissertation.

The initial analysis involved binary logit mode choice models which were estimated across different income segments and household clusters to examine differences in decision making between low, medium, and higher income households and between demographically similar households. The results indicated that the income-segmented models yielded different results than the pooled model at the 95% confidence level, but the parameters were largely consistent across the three segments. The clustered households exhibited more variation in their responses, particularly for the older and larger households. For the year studied, rates of HOT lane use were fairly consistent across the three income groups for which data were available, differing by a maximum of 3.9%. Disaggregate elasticity values revealed low sensitivities to nearly all of the explanatory parameters with the exception of the problematic trip distance variable, and with income among the higher income users. These elasticity values illustrated varying responses to household income and education, for example, across the segmented and clustered households.

The extensions of the preliminary analysis revealed the benefits of further segmenting households by income to illustrate the variety of behavior within the higher income households. This segmentation indicated that the three-segment strategy

disguised substantial behavioral differences among the highest income households on the I-85 corridor. The determinants of lane choice decision-making in the morning peak had notable differences from the determinants of the afternoon peak, particularly with regards to toll rate sensitivity and the impact of the total corridor segments traversed. Afternoon peak models had better goodness of fit metrics overall, though the pseudo-R^2 measures for both time frames were under 0.400 in all but one of the cases. This indicates that the there are many other factors in play in lane choice decision making; the survey and stated preference data that is missing from this analysis may play an important role in improving those models. The operational characteristics included in the lane choice models, including average lane speeds and transponder counts, yielded similar responses across the income segments under examination. It should be noted that the users examined in this study all had registered for Peach Pass transponders, and as such represent a self-selecting sample of corridor users. The similarities in decision making factors across the different models and income groups examined is likely a result of this effect. This issue could begin be addressed by providing transponders automatically and without cost to those users without Peach Pass accounts, though the sample would still be restricted to those users who choose to use them in their vehicles.

The mixed logit framework improved the modeling results by addressing the issue of serial correlation and by estimating the toll amount and household income coefficients as random rather than fixed parameters. The toll amount coefficients, for example, were more appropriately modeled as normal distributions that encapsulate both positive and negative values to reflect both the 'signaling' and demand-reducing effects of toll rates. Further segmenting the households showed that lane choice determinants varied more

within the 'Higher' income segment than across the original three-segment structure. In particular, the five-segment models illustrated lower elasticities with regard to corridor segment counts and toll levels for the highest-income households in the sample, as well as higher household income level elasticities for afternoon trips by that same cohort.

Contributions

This dissertation makes a number of contributions to the study of road pricing in general and High Occupancy Toll lanes in particular. The analysis was among the first in the available literature to use revealed preference lane use data for both the toll lane users and the unpriced general purpose lane users. The use of household level marketing data, rather than census or survey data, was another unique characteristic of this research. Both of these factors involve the application of existing methods to new and unique data sources.

This dissertation outlined the process and pitfalls of combining these two large, unique data sets to generate a new one. The research also provided an overview of the characteristics and shortcomings of these specific lane use and demographic data sets. In addition, this dissertation contrasted the use of marketing data with US Census data to outline the differences and potential biases that resulted from this choice of data sources.

At the time of this writing, the author had not found any investigations of toll lane use that address repeated observations by users. The existing literature did not address serial correlation concerns among HOT lane use; this dissertation used panel data and the mixed logit framework for that purpose. Nor did the author find any examination of the spectrum of responses to factors such as toll amounts; hence this dissertation contributes both a better understanding of the range of possible responses and the differences of those

responses among household income segments. In addition to the use of new types of data sources, this dissertation also provides some of the first applications of more advanced modeling techniques to an area that has not yet seen them.

This research will also provide the basis for a modeling tool that can use the results of this work to investigate Express Lane use decisions in other contexts. The models that form the basis of this dissertation could potentially be generalized to other cities and facilities. The factors included are common enough to allow other researchers to closely replicate their design given similar data availability. These models can be used to better understand the potential rates and factors concerning toll lane use for different demographic groups in other locations. The research involved the development of data processing and modeling scripts that constructed trips from disaggregated vehicle detections, estimated corridor conditions such as travel speeds and travel time reliability, and paired trip records with account, toll, and demographic data to provide a comprehensive overview of user characteristics and operating conditions at all times under examination.

These results have implications for both existing and new priced facilities. Along with previous work by Smith (2011) and Khoeini (2014), this research offers evidence regarding the 'Lexus Lane' moniker applied to High Occupancy Toll facilities. Rates of priced lane use and the determinants of lane choice decisions were consistent across income segments, with the exception of the highest income households in the sample. The analysis of value of travel time savings with a demographic component that looks at household income has not yet been seen in the literature; similarly, the findings regarding differing behavior among very high income households appear to be unseen in the

existing literature. The use of new data sources, the evaluation of those types of data sources, and the application of methods that have previously been unused in this field make up the primary contributions of this dissertation.

Limitations of study

The data set used in this research was rich and innovative. The Express Lanes data streams provided tremendous amounts of very detailed records in both the HOT and General Purpose Lanes, including unique data with decisions to use and not use the toll lanes. The Epsilon credit report data allowed this dissertation to examine recent socioeconomic characteristics for a large sample of households. While the data had these advantages, among others, they also came with their own issues.

Match Rates and Sample Bias in Study Data

The analytical process revealed a number of ways in which the study data may have been biased. The mechanisms that created the possibility for this bias included matching the SRTA data with the vehicle registration database, matching those results with the demographic data, and constructing the complete data set. The impacts of these processing stages were seen in the subset of Peach Pass transponders and Epsilon households that were present in the final data set. The resulting sample differed from the complete set of SRTA data by primarily including those vehicles that frequently used the corridor; the bottom quartile of users ranked by trip frequency were virtually excluded from the paired sample.

The effects of the various data processing steps in this dissertation on the demographics of the sample were examined in different ways. The Connecting SRTA Data to Epsilon Data chapter compared the paired demographic data with the full data

purchase. That chapter also compared the paired households with City of Atlanta dwellers using Census ACS data. The Potential Sample Bias in Paired Vehicle Activity and Marketing Data chapter examined the ACS-provided demographic characteristics of the GRTI-matched households, Epsilon households, and the households for which the SRTA-Epsilon pairing was successful. That investigation found a substantial bias in the SRTA-Epsilon paired sample towards higher income households, while the other demographic characteristics examined were largely similar.

This dissertation also examined the data loss that occurred in joining the SRTA constructed trips with the Epsilon demographic data and with the other streams and data sets that were originated with SRTA. The joining process resulted in the exclusion of a significant portion of the constructed trip population: the trips that remained at the end of the process differ primarily in the higher rates of toll lane use, lower average speeds, and fewer households represented. The lack of a general purpose lane reader on SR-316 meant that trips that started or ending on that corridor segment were excluded from the final analytical dataset. As those trips represent roughly a quarter of southbound morning peak-period trips, the loss of data is substantial. The structure of the Account data stream was another potential source of bias: left unaddressed, the many-to-many relationships in the data stream can restrict analysis only to those accounts with a single transponder and vehicle. Finally, the registration database used for the license plate matching may not reflect the actual garage location of the vehicles. That is, the registration address may differ from the current address (Nelson et al., 2008).

Revealed Preference Data

The data used in this proposed study was strictly of the revealed preference variety. Other work of this type often includes stated preference components, such as survey results, to fill in the holes left by revealed preference data (Bhat & Castelar, 2002; Borjesson, 2006). Without a survey component, this study could not separate trips by trip purpose, for example. This would be a useful form of segmentation in the choice models as commute trips may inspire different behavior than leisure or shopping trips. Additionally, trip purpose is typically identified as a significant determinant in studies that examine willingness-to-pay (Jiang, 2004). Other potentially useful characteristics that may be provided by a survey include job type, school location, use of day care, and more. Similarly, relying only on revealed preference data means that this dissertation included no information on the schedules of the drivers. Schedule preference information may be useful in choice studies as users may behave differently if they are late to their destination, such as work, versus if they are on time or early. Other research assigns different utility impacts to early or late arrivals; without such information, this study cannot make such a distinction. Additional missing elements with potential model impacts that could be provided by household surveys include trip start time versus work start time, day care or sports attendance, and trip purpose and destination. Survey data would also be useful in confirming the household characteristics provided by the privately sourced demographic data. Finally, stated preference data allows researchers to measure user perceptions rather than actual behavior. In this context, user perceptions of travel time, travel time savings, and travel time reliability would all be relevant to the analysis. Travel time savings, for example, are typically perceived to be much higher

than they are in actuality (Devarasetty & Burris, 2013). The perceptions of these measures may explain HOT lane choice better than the actual time savings or reliability values; the models may suffer for lack of them. Capturing the history of a user's experience along the corridor may have similar effects; an investigation of the literature concerning the connection between history and perception is still ongoing.

Data Limitations

One of the most unique features of the data set was the availability of Peach Pass tag reads in the General Purpose lanes, which allowed the author to examine when tag holders do not use the HOT lanes. Unfortunately, one major section of the corridor was missing a General Purpose Peach Pass scanner. State Route 316, which contains a branch of the Express Lanes, does not have a GP scanner and so could not be used in the direct travel time comparisons. Figure 165 shows the segment of the Express Lanes corridor that includes SR 316; the yellow bars ("G-35") indicate HOT gantries, while the green bars ("SCAN-N6") represent GP lane gantries. In the southbound direction, 23.7% (482171/2033104) of all of the trips in 2012 began on SR 316, and 15.9% (321693/2021543) of northbound 2012 trips ended on SR316. These trips were excluded from the analyses, as the data did not allow for travel time comparisons for that segment of the corridor.

Figure 165: Express Lanes Diagram - SR316 Focus (Source: Atkins I-85 Express Lane Straight Line Diagram)

Another limiting factor of the study also concerned the geographical reach of the data: it is limited to the I-85 corridor. That is, the travel times and lane choice decisions and conditions within the scope of the data did not encompass the entire trips made by the users. The trips represented in the data were incomplete, and this may have impacted the results. For example, the decision to use the Express Lanes may be based on total trip cost; users may be more willing to pay the toll if it is a smaller proportion of their total cost (Li, 2001). Similarly, the total trip distance may make a user more or less likely to purchase better service for a portion of that trip. These effects could not be estimated given the available data. This is in addition to the trip factors identified above as potentially available in survey data: trip purpose, work start time, trip destination, etc.

Future Work

Despite the enhancements made to the preliminary models in this chapter, further model development remains both possible and desirable. While the set of variables used in this analysis greatly expanded upon those used in the preliminary modeling efforts, further complications arose in the form of collinearity among the operational factors. The transponder count factor was included to provide some measure of the overall demand, though a more comprehensive measure that was not restricted to Peach Pass-holding vehicles would be preferred. Additional exploration of the five-segment models with the mixed logit framework, while time consuming, would likely provide more insights. For the purpose of designing a modeling tool that could be used in other locations, it would be beneficial to perform validation tests on the selected models to investigate their accuracy. While this analysis compared lane use behavior across income segments, further work could be done in examining the overall welfare benefits of the facility for users and non-users. The data loss that occurred as part of the dataset construction process could be addressed with imputation methods that should reduce the bias caused by unsuccessful database joins. Improving the match rates between the SRTA and Epsilon data and addressing the resulting sample bias that occurs, outlined earlier in this dissertation, could improve models and provide a more comprehensive overview of the users and non-users of the Express Lanes.

The data set used and described in this dissertation is rich and very large in its scope, with ample opportunity for additional analyses. The marketing data elements that were used in this dissertation ultimately comprised a small subset of the available data; further explorations of household demographics such as occupation and retirement status

could improve the models and behavioral understanding. The value of travel time savings investigation revealed large differences among trips of differing origins, destinations, and lengths; a more thorough examination of this phenomenon would be worthwhile. Toll lane use was examined in a limited binary fashion; much remains to be done in examining differing lengths of toll lane trips and the decisions made at each potential weave zone. A closer look at the data loss caused by the various joins in the data set construction could reduce potential bias. Further comparisons between the marketing data and the census data, particularly with regards to their impact on the modeling work, would likely be beneficial. New data could also provide many ways to expand on this work. In particular, an expanded household demographic data set could increase the sample of households and transponders in the analysis. Finally, survey data that provides stated preference and other data would be a valuable way to supplement the analyses conducted here and to give a more comprehensive view of user behavior.

APPENDIX A

CORRELATION MATRICES

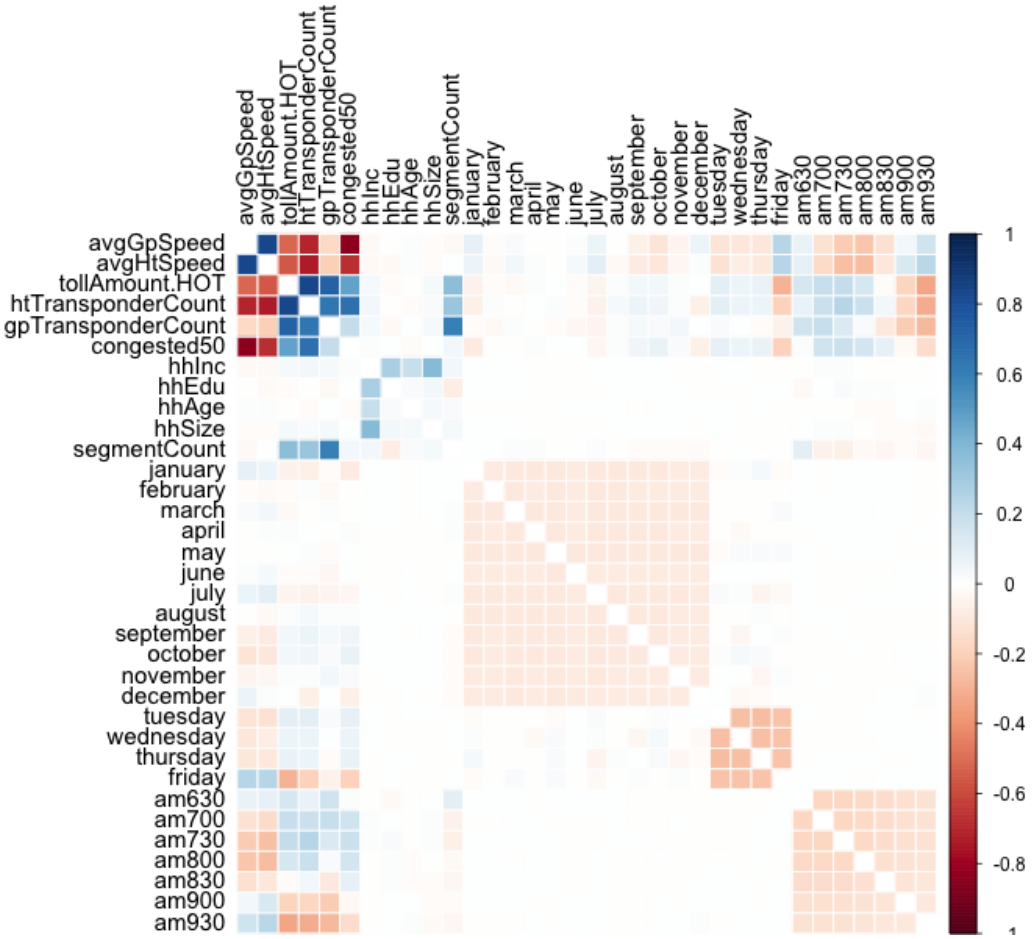


Figure 166: AM Peak Period Trips – Model 9 Variable Correlation Matrix

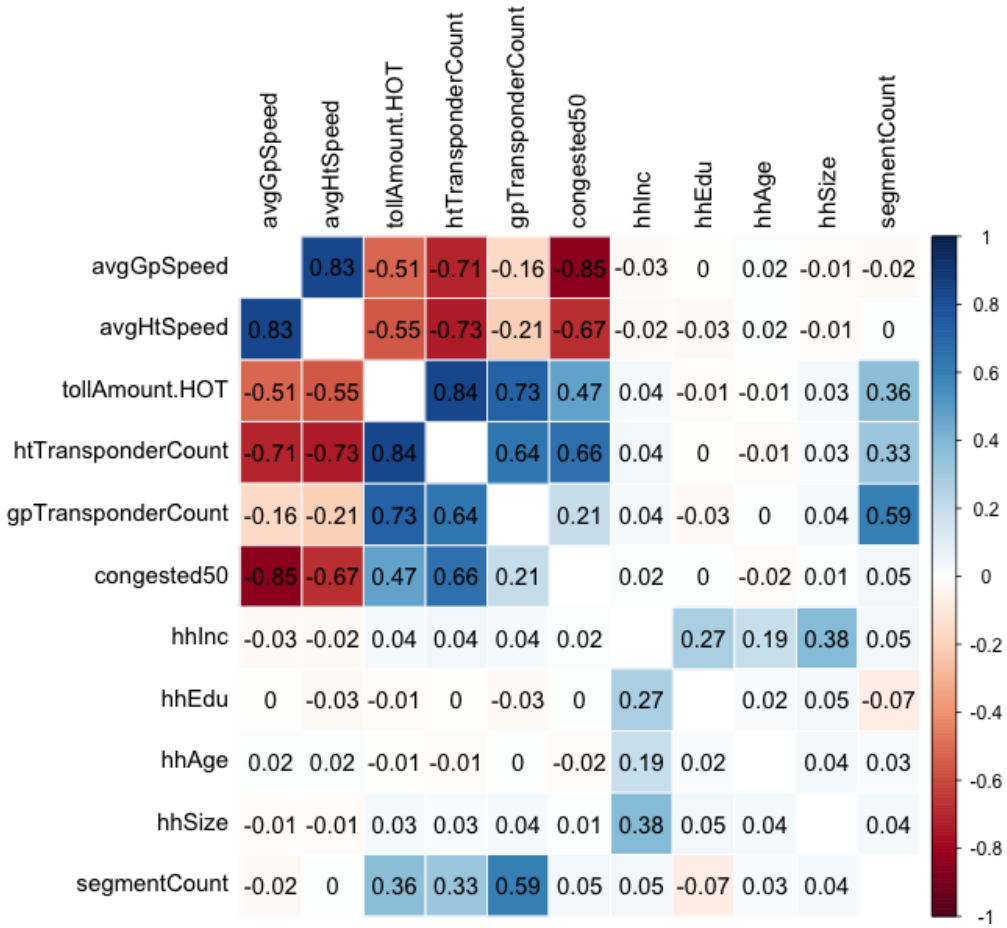


Figure 167: AM Peak Period Trips – Model 9 Variables Minus Time/Date Indicators Correlation Matrix

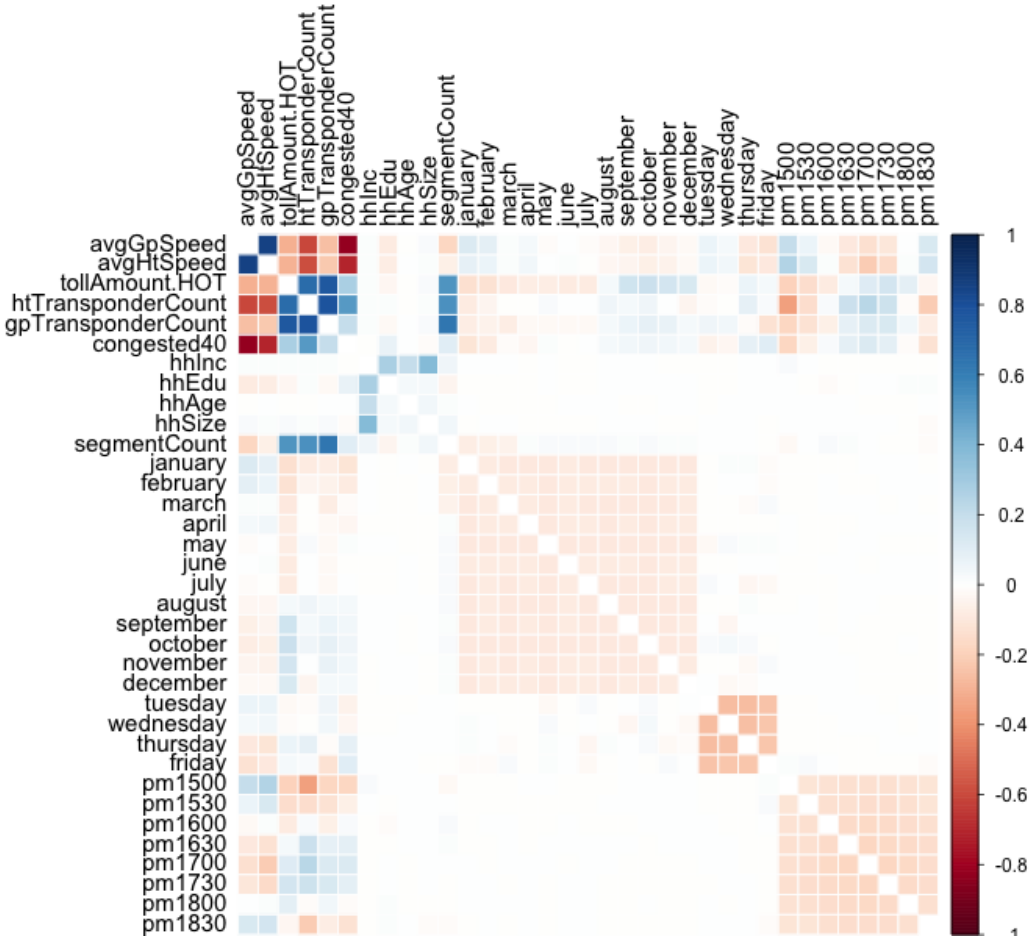


Figure 168: PM Peak Period Trips – Model 9 Variables - Correlation Matrix

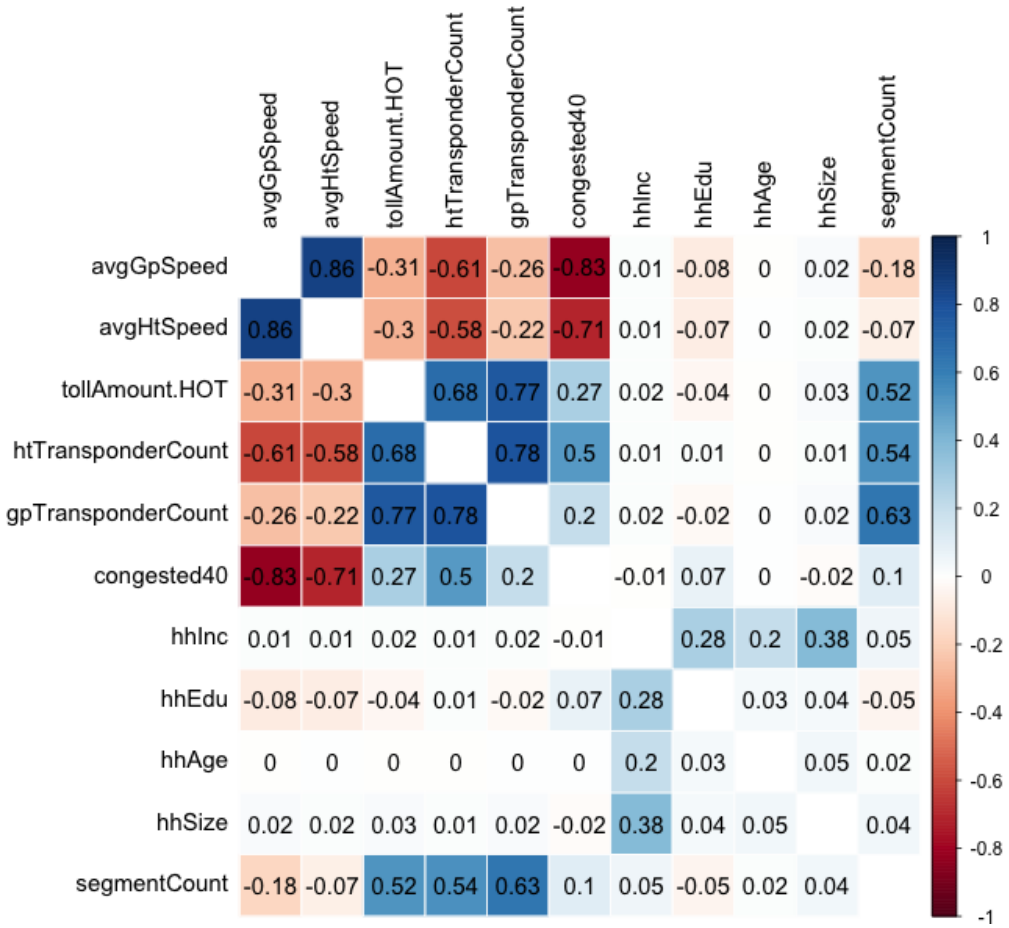


Figure 169: PM Peak Period Trips – Model 9 Variables Minus Time/Date Indicators - Correlation Matrix

APPENDIX B

FITTING VALUE OF TRAVEL TIME SAVINGS DISTRIBUTIONS

The distributions for the value of travel time savings exhibited by the users of the I-85 Express Lanes were relatively consistent in shape across different years and income segments, though the precise measures of centrality and dispersion differed. In particular, the value of travel time savings distributions resemble a gamma distribution. To try and recreate these shapes consistently, researchers sought to fit the data to different distributional curves. This section shows the results of this distribution-fitting analysis for the value of travel time savings distributions of each income segment in the southbound AM peak period and the northbound PM peak period. In addition to the gamma distribution, researchers fit the exponential, logistic, and Weibull distributions as well.

For each value of travel time savings distribution, researchers estimated distributional parameters for the four distributions named above. The author then drew 100,000 random draws from each distribution using the parameters that were fit to the actual data. This section presents the results of those estimates and the resulting curves. The original distribution is presented in blue, while the attempts to fit the four distributions are transparently overlaid on top. Table 113 and Table 114 summarize the parameters for each category of distribution and provide the resulting test statistic from the Kolmogorov-Smirnov test. Researchers employed this test to investigate the suitability of the fitted distributions; the null hypothesis of this test says that the two samples are drawn from the same distribution. For each case presented here, the test compares the original VTTS data with the fitted distributions.

Fitting Southbound VTTS Data to Distributions

Figure 170 presents the southbound AM peak period value of travel time savings distribution for calendar year 2013 along with the fitted distribution curves. The top-most chart shows the results from the lower income segment. The gamma, Weibull, and exponential curves are similar to each other, though a visual inspection suggests that none resemble the original distribution very closely. The logistic curve differs the most from the others, with a shifted center and a higher concentration of values between roughly $50/hour and $100/hour. Because the maximum VTTS value in each segment exceeded $1000/hour, each curve was fitted to the subset of VTTS data in which the value of travel time savings was greater than zero and less than $300/hour.

The second chart in Figure 170 shows the medium income segment distribution and the corresponding estimated curve results. The distributions all resemble the lower income results very closely; the summary provided in Table 113 below shows how close they are. Just as in the lower income segment results, none of the resulting curves appear to fit the original data very closely. The final plot provides the results for the higher income, southbound 2013 distribution. Once again, the curves all resemble their counterparts from the lower and middle income segments very closely, and none of them appear to fit the original data very well. In each case, the gamma and weibull distributions most closely approximate the location of the peak of the distribution but not the peak itself. The exponential curve better models the peak of the actual data, but here the location is less accurate. The logistic distribution curve does not resemble the shape of the original distribution in either aspect.

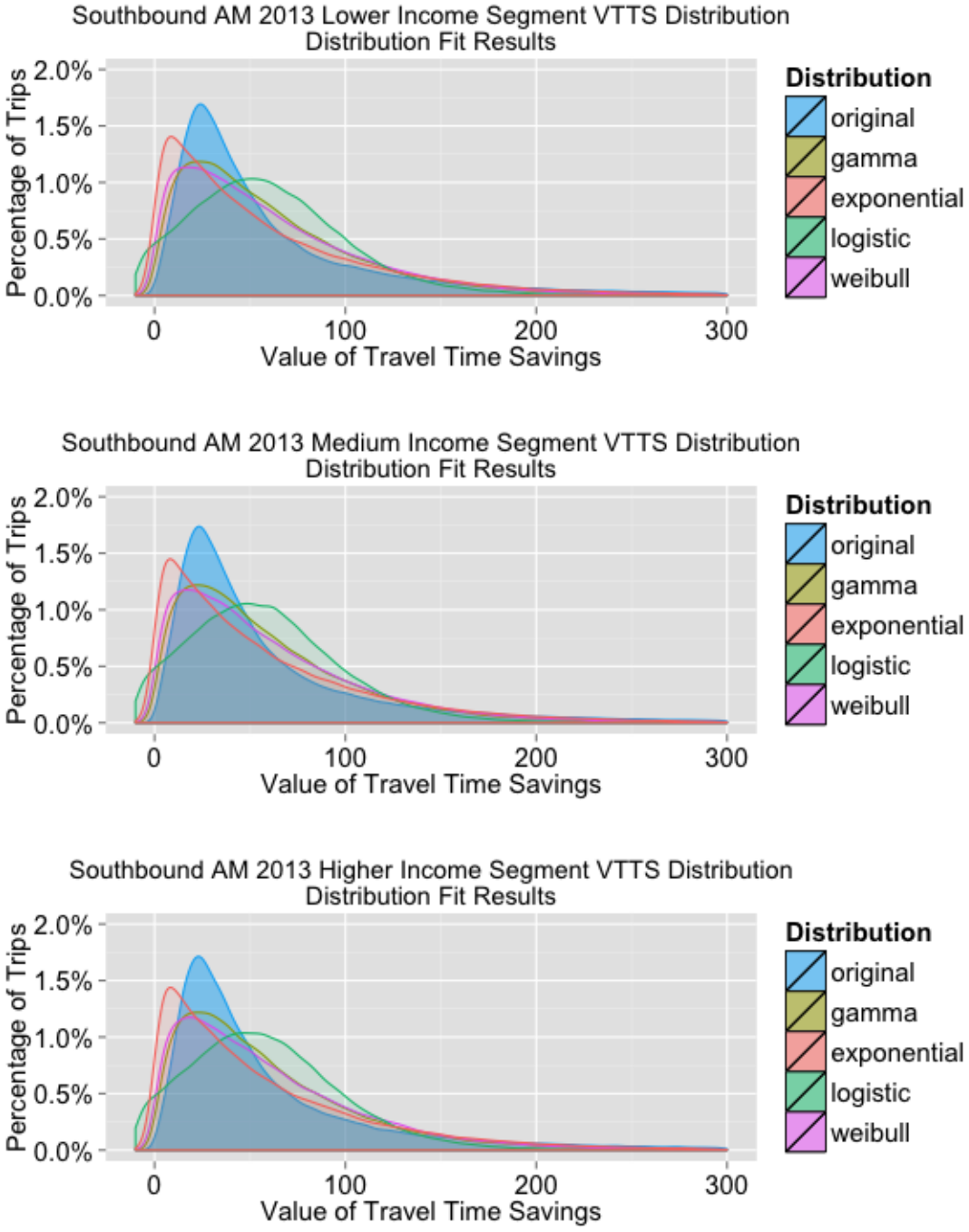


Figure 170: Southbound VTTS Distribution Fit Curves

Table 113 presents the estimated parameters for each of the four distribution categories along with the Kolmogorov-Smirnov test p-values. As one might expect from the visual similarity of all of the estimated curves presented above, the estimated parameters for each distribution are nearly identical. Of the four fitted distributions, the logistic parameters vary the most across the three income segments. The Kolmogorov-Smirnov test results are consistent for each distribution type and income segment: in every case, the null hypothesis of equal distributions is rejected at well over the 99% confidence level.

Table 113: Southbound VTTS Distribution Fit Results

	Lower Income	Medium Income	Higher Income
Gamma Distribution			
Shape	1.57	1.56	1.58
Rate	0.026	0.026	0.027
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Weibull Distribution			
Shape	1.23	1.22	1.23
Scale	64.99	63.19	63.81
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Logistic Distribution			
Location	50.81	49.38	49.87
Scale	26.73	26.11	26.19
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Exponential Distribution			
Rate	0.017	0.017	0.017
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$

Fitting Northbound VTTS Data to Distributions

The results of the distribution fitting analysis for the northbound, PM peak period trips in 2013 are shown in Figure 171. As in the previous charts for the southbound trips, the results for the three income segments are very similar. Once again, the gamma and (to a lesser extent) Weibull distributions approximate the location of the distributional peak, while the exponential distribution better approximates the magnitude of said peak. Because the northbound VTTS distributions are narrower, the data subset for distribution fitting was restricted to VTTS values between $0/hour and $150/hour. Despite this narrower estimation frame, none of the estimated distributions resembled the original data well.

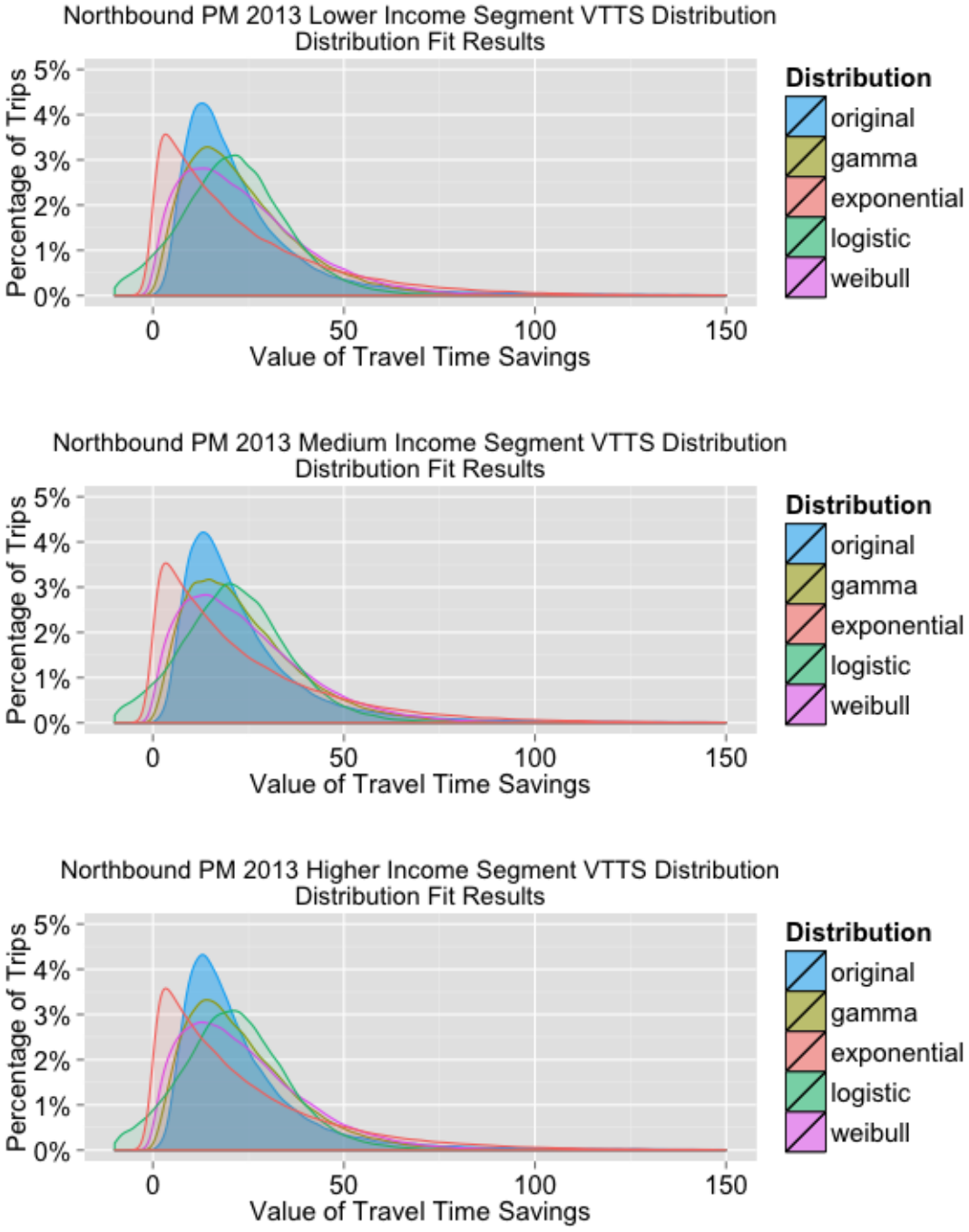


Figure 171: Northbound VTTS Distribution Fit Curves

Table 114 presents the parameters of the fitted distributions for the northbound PM peak period trips in 2013. As in the southbound results, the parameters for each distribution type are similar across income segments. Again, the Kolmogorov-Smirnov test for distributional equality results in the rejection of the null hypothesis in each case.

Table 114: Northbound VTTS Distribution Fit Results

	Lower Income	Medium Income	Higher Income
Gamma Distribution			
Shape	2.44	2.47	2.50
Rate	0.10	0.10	0.11
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Weibull Distribution			
Shape	1.49	1.50	1.50
Scale	26.33	26.66	26.31
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Logistic Distribution			
Location	20.73	21.01	20.71
Scale	8.34	8.40	8.23
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$
Exponential Distribution			
Rate	0.042	0.042	0.043
Kolmogorov-Smirnov Test Result	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$	$p<2.2x10^{-16}$

Distributional Fitting Overview

The purpose of fitting distributions to the value of travel time savings data was to investigate whether this behavior of the I-85 Express Lanes users could be replicated for future projects using a standard distribution. That is, whether the value of travel time savings distributions could be recreated using an equation containing one or two estimated parameters. The results indicated that the shapes of the actual VTTS data are difficult to recreate with the four types of distributions examined here. Through visual inspection and the Kolmogorov-Smirnov test, researchers saw that the unique shape of

the actual VTTS data does not lend itself to being approximated by standard distributions. It should be noted that the shapes of the estimated distributions were sensitive to the subset of actual data over which they were estimated; narrowing the range of VTTS results from a maximum of $500/hour to $300/hour or $150/hour made the fitted curves more closely resemble the actual distributions. This had limited impact, however, as even the more restricted data ranges did not produce accurate fits among the estimated curves.

APPENDIX C

ODDS RATIOS

Table 115: Model 1 Odds Ratios

	AM Peak – Model 1	PM Peak – Model 1
Intercept	0.107	0.023
avgSpeed	0.971	1.02
tollAmount	0.572	0.573
transponderCount	1.00	1.00
HOT: congested50	4.35	
HOT: congested40		3.93
HOT: log(hhIncomedollars)	1.04	1.06
HOT: hhEdu	1.01	0.938
HOT: hhAge	0.970	0.987
HOT: hhSize	0.950	0.972
HOT: distancemi	1.44	1.54

Table 116: Model 2 Odds Ratios

	AM Peak – Model 2	PM Peak – Model 2
Intercept	0.076	0.020
avgSpeed	0.990	1.01
tollAmount	0.704	0.711
transponderCount	1.00	1.00
HOT: congested50	4.03	
HOT: congested40		3.29
HOT: log(hhIncomedollars)	1.13	1.14
HOT: hhEdu	0.908	0.818
HOT: hhAge	0.972	0.994
HOT: hhSize	0.957	0.980
HOT: segmentCount	1.93	2.60

Table 117: Model 3 Odds Ratios

	AM Peak – Model 3	PM Peak – Model 3
Intercept	0.088	0.0313
$avgSpeed^2$	1.00	1.00
tollAmount	0.691	0.696
transponderCount	1.00	1.00
HOT: congested50	4.17	
HOT: congested40		3.95
HOT: log(hhIncomedollars)	1.14	1.13
HOT: hhEdu	0.902	0.823
HOT: hhAge	0.971	0.994
HOT: hhSize	0.956	0.979
HOT: segmentCount	1.96	2.71

Table 118: Model 5 Odds Ratios

	AM Peak – Model 5	PM Peak – Model 5
Intercept	0.0666	0.0313
$avgSpeed^2$	1.00	1.00
tollAmount	0.647	0.696
transponderCount	1.00	1.00
HOT: congested50	7.80	
HOT: congested40		3.95
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.889	0.823
HOT: hhAge	0.970	0.994
HOT: hhSize	0.957	0.979
HOT: segmentCount	2.00	2.71

Table 119: Model 6 Odds Ratios

	AM Peak – Model 6	PM Peak – Model 6
Intercept	0.0601	0.0312
avgSpeed^2	1.00	1.00
tollAmount	0.646	0.655
transponderCount	1.00	1.00
HOT: congested50	7.71	
HOT: congested40		3.93
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.889	0.820
HOT: hhAge	0.970	0.994
HOT: hhSize	0.957	0.979
HOT: segmentCount	2.00	2.87
HOT: february	1.07	0.943
HOT: march	1.08	0.957
HOT: april	1.17	0.773
HOT: may	1.15	0.699
HOT: june	1.10	0.705
HOT: july	1.10	0.720
HOT: august	1.16	0.849
HOT: september	1.21	1.21
HOT: october	1.24	1.25
HOT: november	1.13	1.17
HOT: december	1.02	1.15

Table 120: Model 6b Odds Ratios

	AM Peak – Model 6b	PM Peak – Model 6b
Intercept	0.0619	0.0330
avgSpeed^2	1.00	1.00
tollAmount	0.646	0.664
transponderCount	1.00	1.00
HOT: congested50	7.71	
HOT: congested40		3.93
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.889	0.820
HOT: hhAge	0.970	0.994
HOT: hhSize	0.957	0.979
HOT: segmentCount	2.00	2.84
HOT: spring	1.10	0.784
HOT: summer	1.09	0.741
HOT: fall	1.16	1.16

Table 121: Model 7 Odds Ratios

	AM Peak – Model 7	PM Peak – Model 7
Intercept	0.0584	0.0322
$avgSpeed^2$	1.00	1.00
tollAmount	0.619	0.658
transponderCount	1.00	1.00
HOT: congested50	7.80	3.92
HOT: congested40		
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.891	0.820
HOT: hhAge	0.968	0.994
HOT: hhSize	0.958	0.979
HOT: segmentCount	2.06	2.87
HOT: february	1.11	0.940
HOT: march	1.13	0.952
HOT: april	1.23	0.770
HOT: may	1.21	0.698
HOT: june	1.13	0.703
HOT: july	1.10	0.713
HOT: august	1.21	0.845
HOT: september	1.29	1.20
HOT: october	1.29	1.25
HOT: november	1.20	1.17
HOT: december	1.06	1.15
HOT: tuesday	1.14	1.02
HOT: wednesday	1.18	1.03
HOT: thursday	1.17	0.915
HOT: friday	0.621	0.937

Table 122: Model 8 Odds Ratios

	AM Peak – Model 8	PM Peak – Model 8
Intercept	0.0356	0.0270
$avgSpeed^2$	1.00	1.00
tollAmount	0.577	0.636
transponderCount	1.00	1.00
HOT: congested50	6.28	
HOT: congested40		3.93
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.883	0.812
HOT: hhAge	0.970	0.994
HOT: hhSize	0.956	0.980
HOT: segmentCount	2.25	2.98
HOT: february	1.19	0.932
HOT: march	1.17	0.944
HOT: april	1.30	0.761
HOT: may	1.30	0.689
HOT: june	1.20	0.690
HOT: july	1.12	0.698
HOT: august	1.32	0.855
HOT: september	1.47	1.26
HOT: october	1.46	1.31
HOT: november	1.32	1.24
HOT: december	1.13	1.22
HOT: tuesday	1.19	1.02
HOT: wednesday	1.21	1.03
HOT: thursday	1.21	0.898
HOT: friday	0.53	0.918
HOT: sevenAm	2.19	
HOT: eightAm	1.88	
HOT: nineAm	1.03	
HOT: fourPm		0.955
HOT: fivePm		1.18
HOT: sixPm		1.49

Table 123: Model 9 Odds Ratios

	AM Peak – Model 9	PM Peak – Model 9
Intercept	0.0137	0.0273
$avgSpeed^2$	1.00	1.00
tollAmount	0.508	0.632
transponderCount	1.00	1.00
HOT: congested50	4.95	
HOT: congested40		3.92
HOT: log(hhIncomedollars)	1.15	1.13
HOT: hhEdu	0.877	0.811
HOT: hhAge	0.973	0.995
HOT: hhSize	0.958	0.980
HOT: segmentCount	2.53	3.00
HOT: february	1.30	0.931
HOT: march	1.26	0.944
HOT: april	1.46	0.760
HOT: may	1.46	0.688
HOT: june	1.32	0.689
HOT: july	1.16	0.698
HOT: august	1.51	0.858
HOT: september	1.80	1.28
HOT: october	1.79	1.33
HOT: november	1.57	1.25
HOT: december	1.24	1.23
HOT: tuesday	1.27	1.02
HOT: wednesday	1.27	1.03
HOT: thursday	1.26	0.898
HOT: friday	0.41	0.918
HOT: am630	4.90	
HOT: am700	6.91	
HOT: am730	7.30	
HOT: am800	6.30	
HOT: am830	4.92	
HOT: am900	3.07	
HOT: am930	1.44	
HOT: pm1530		0.947
HOT: pm1600		0.873
HOT: pm1630		0.986
HOT: pm1700		1.08
HOT: pm1730		1.23
HOT: pm1800		1.50
HOT: pm1830		1.41

Table 124: Model 10 Odds Ratios

	AM Peak – Model 10	PM Peak – Model 10
Intercept	0.0118	0.0376
avgSpeed^2	1.00	1.00
tollAmount^2	0.936	0.954
transponderCount	0.999	1.00
HOT: congested50	3.84	
HOT: congested40		3.43
HOT: log(hhIncomedollars)	1.14	1.13
HOT: hhEdu	0.879	0.815
HOT: hhAge	0.974	0.995
HOT: hhSize	0.958	0.980
HOT: segmentCount	2.14	2.60
HOT: february	1.25	0.917
HOT: march	1.21	0.897
HOT: april	1.34	0.736
HOT: may	1.37	0.677
HOT: june	1.27	0.687
HOT: july	1.13	0.689
HOT: august	1.40	0.763
HOT: september	1.69	1.05
HOT: october	1.68	1.08
HOT: november	1.49	1.05
HOT: december	1.18	1.03
HOT: tuesday	1.27	1.00
HOT: wednesday	1.28	0.996
HOT: thursday	1.27	0.834
HOT: friday	0.538	0.844
HOT: am630	3.46	
HOT: am700	4.87	
HOT: am730	5.12	
HOT: am800	4.53	
HOT: am830	3.86	
HOT: am900	3.07	
HOT: am930	2.20	
HOT: pm1530		0.867
HOT: pm1600		0.737
HOT: pm1630		0.766
HOT: pm1700		0.808
HOT: pm1730		0.905
HOT: pm1800		1.13
HOT: pm1830		1.16

Table 125: Model 11 Odds Ratios

	AM Peak – Model 11	PM Peak – Model 11
Intercept	0.0271	0.00776
avgSpeed^2	1.00	1.00
tollAmount	0.488	0.614
htDensity	0.996	1.01
HOT: congested50	5.12	
HOT: congested40		4.45
HOT: log(hhIncomedollars)	1.14	1.14
HOT: hhEdu	0.880	0.808
HOT: hhAge	0.973	0.993
HOT: hhSize	0.958	0.980
HOT: segmentCount	2.41	3.19
HOT: february	1.30	0.963
HOT: march	1.28	1.05
HOT: april	1.49	0.771
HOT: may	1.48	0.703
HOT: june	1.32	0.718
HOT: july	1.16	0.731
HOT: august	1.54	0.839
HOT: september	1.87	1.13
HOT: october	1.87	1.14
HOT: november	1.65	1.02
HOT: december	1.27	1.01
HOT: tuesday	1.26	1.07
HOT: wednesday	1.26	1.11
HOT: thursday	1.25	1.13
HOT: friday	0.387	1.30
HOT: am630	5.41	
HOT: am700	7.78	
HOT: am730	8.14	
HOT: am800	6.71	
HOT: am830	4.99	
HOT: am900	2.94	
HOT: am930	1.31	
HOT: pm1530		1.16
HOT: pm1600		1.16
HOT: pm1630		1.25
HOT: pm1700		1.32
HOT: pm1730		1.41
HOT: pm1800		1.61
HOT: pm1830		1.44

Table 126: Model 12 AM Peak Odds Ratios

	AM Peak – Model 12a	AM Peak – Model 12b
Intercept	0.0751	0.134
avgSpeed^2	1.00	1.00
Toll/log(income)	0.000647	0.000617
transponderCount	1.00	1.00
HOT: congested50	4.90	4.91
HOT: log(hhIncomedollars)		0.937
HOT: hhEdu	0.866	0.880
HOT: hhAge	0.967	0.973
HOT: hhSize	0.952	0.960
HOT: segmentCount	2.50	2.52
HOT: february	1.30	1.30
HOT: march	1.26	1.26
HOT: april	1.46	1.46
HOT: may	1.45	1.45
HOT: june	1.31	1.31
HOT: july	1.16	1.16
HOT: august	1.50	1.50
HOT: september	1.78	1.78
HOT: october	1.77	1.78
HOT: november	1.55	1.56
HOT: december	1.23	1.23
HOT: tuesday	1.27	1.27
HOT: wednesday	1.28	1.27
HOT: thursday	1.27	1.26
HOT: friday	0.421	0.418
HOT: am630	4.77	4.80
HOT: am700	6.70	6.76
HOT: am730	7.08	7.15
HOT: am800	6.15	6.19
HOT: am830	4.84	4.86
HOT: am900	3.05	3.06
HOT: am930	1.45	1.45

Table 127: Model 12 PM Peak Odds Ratios

	PM Peak – Model 12a	PM Peak – Model 12b
Intercept	0.101	0.0713
avgSpeed^2	1.00	1.00
Toll/log(income)	0.00643	0.00658
transponderCount	1.00	1.00
HOT: congested40	3.91	3.91
HOT: log(hhIncomedollars)		1.04
HOT: hhEdu	0.821	0.813
HOT: hhAge	0.999	0.995
HOT: hhSize	0.985	0.981
HOT: segmentCount	3.00	2.99
HOT: february	0.931	0.931
HOT: march	0.943	0.943
HOT: april	0.759	0.759
HOT: may	0.687	0.688
HOT: june	0.688	0.689
HOT: july	0.697	0.697
HOT: august	0.856	0.854
HOT: september	1.27	1.27
HOT: october	1.32	1.32
HOT: november	1.25	1.24
HOT: december	1.22	1.22
HOT: tuesday	1.02	1.02
HOT: wednesday	1.03	1.03
HOT: thursday	0.897	0.896
HOT: friday	0.916	0.915
HOT: pm1530	0.944	0.943
HOT: pm1600	0.869	0.868
HOT: pm1630	0.981	0.979
HOT: pm1700	1.07	1.07
HOT: pm1730	1.22	1.22
HOT: pm1800	1.49	1.49
HOT: pm1830	1.41	1.40

Table 128: Model 13 AM Peak Odds Ratios

	AM Peak – Model 13a	AM Peak – Model 13b
Intercept	0.176	1.38
avgSpeed^2	1.00	1.00
Toll/income	0.000	0.000
transponderCount	0.999	0.999
HOT: congested50	2.70	2.75
HOT: log(hhIncomedollars)		0.807
HOT: hhEdu	0.869	0.899
HOT: hhAge	0.971	0.984
HOT: hhSize	0.953	0.971
HOT: segmentCount	1.55	1.59
HOT: february	1.18	1.18
HOT: march	1.16	1.16
HOT: april	1.18	1.19
HOT: may	1.18	1.18
HOT: june	1.18	1.18
HOT: july	1.13	1.12
HOT: august	1.20	1.21
HOT: september	1.27	1.28
HOT: october	1.26	1.27
HOT: november	1.13	1.14
HOT: december	0.928	0.933
HOT: tuesday	1.22	1.22
HOT: wednesday	1.27	1.26
HOT: thursday	1.25	1.25
HOT: friday	0.915	0.891
HOT: am630	1.73	1.78
HOT: am700	2.06	2.14
HOT: am730	2.14	2.22
HOT: am800	2.24	2.31
HOT: am830	2.53	2.58
HOT: am900	2.57	2.59
HOT: am930	2.12	2.10

Table 129: Model 13 PM Peak Odds Ratios

	PM Peak – Model 13a	PM Peak – Model 13b
Intercept	0.264	0.625
avgSpeed^2	1.00	1.00
Toll/income	0.000	0.000
transponderCount	1.00	1.00
HOT: congested40	2.89	2.91
HOT: log(hhIncomedollars)		0.912
HOT: hhEdu	0.825	0.838
HOT: hhAge	0.996	1.00
HOT: hhSize	0.980	0.988
HOT: segmentCount	2.36	2.38
HOT: february	0.908	0.909
HOT: march	0.864	0.866
HOT: april	0.730	0.730
HOT: may	0.674	0.674
HOT: june	0.694	0.693
HOT: july	0.684	0.683
HOT: august	0.673	0.677
HOT: september	0.809	0.819
HOT: october	0.822	0.832
HOT: november	0.829	0.837
HOT: december	0.841	0.848
HOT: tuesday	0.981	0.982
HOT: wednesday	0.958	0.960
HOT: thursday	0.749	0.754
HOT: friday	0.749	0.754
HOT: pm1530	0.803	0.805
HOT: pm1600	0.637	0.642
HOT: pm1630	0.604	0.611
HOT: pm1700	0.597	0.607
HOT: pm1730	0.652	0.663
HOT: pm1800	0.856	0.868
HOT: pm1830	1.01	1.02

Table 130: Model 14 AM Peak Odds Ratios

	AM Peak – Model 14a	AM Peak – Model 14b	AM Peak – Model 14c	AM Peak – Model 14d
Intercept	0.0486	0.288	0.0387	0.0107
$avgSpeed^2$	1.00	1.00	1.00	1.00
tollAmount	0.507	0.503	0.509	0.508
transponderCount	1.00	1.00	1.00	1.00
HOT: congested50	4.96	4.34	4.93	4.95
HOT: hhEdu	0.861	0.880	0.906	0.875
HOT: hhAge	0.966	0.969	0.988	0.975
HOT: income/hhSize	1.00	1.00		
HOT: income		0.806		1.15
HOT: hhSize		1.05		0.993
HOT: logIncome/hhSize			1.04	1.05
HOT: segmentCount	2.53	2.61	2.54	2.53
HOT: february	1.30	1.30	1.30	1.30
HOT: march	1.27	1.28	1.26	1.26
HOT: april	1.46	1.49	1.46	1.46
HOT: may	1.46	1.47	1.46	1.46
HOT: june	1.32	1.33	1.31	1.32
HOT: july	1.16	1.16	1.16	1.16
HOT: august	1.51	1.52	1.51	1.51
HOT: september	1.80	1.84	1.80	1.80
HOT: october	1.80	1.87	1.79	1.79
HOT: november	1.57	1.64	1.57	1.57
HOT: december	1.24	1.27	1.24	1.24
HOT: tuesday	1.27	1.27	1.27	1.27
HOT: wednesday	1.27	1.26	1.27	1.27
HOT: thursday	1.27	1.25	1.26	1.26
HOT: friday	0.412	0.402	0.413	0.413
HOT: am630	4.92	4.92	4.89	4.90
HOT: am700	6.93	7.08	6.90	6.90
HOT: am730	7.32	7.54	7.29	7.29
HOT: am800	6.32	6.48	6.27	6.30
HOT: am830	4.92	4.96	4.90	4.92
HOT: am900	3.06	2.98	3.06	3.07
HOT: am930	1.43	1.35	1.44	1.44

Table 131: Model 14 PM Peak Odds Ratios

	PM Peak – Model 14a	PM Peak – Model 14b	PM Peak – Model 14c	PM Peak – Model 14d
Intercept	0.0868	0.135	0.0801	0.0250
$avgSpeed^2$	1.00	1.00	1.00	1.00
tollAmount	0.632	0.633	0.633	0.632
transponderCount	1.00	1.00	1.00	1.00
HOT: congested40	3.92	3.90	3.90	3.92
HOT: hhEdu	0.810	0.807	0.837	0.811
HOT: hhAge	0.994	0.992	1.01	0.995
HOT: income/hhSize	1.00	1.00		
HOT: income		0.947		
HOT: hhSize		1.03		
HOT: logIncome/hhSize			1.01	1.02
HOT: segmentCount	3.00	2.99	3.01	3.00
HOT: february	0.931	0.930	0.931	0.931
HOT: march	0.944	0.942	0.944	0.944
HOT: april	0.760	0.757	0.758	0.760
HOT: may	0.688	0.687	0.687	0.688
HOT: june	0.690	0.690	0.689	0.689
HOT: july	0.699	0.697	0.758	0.698
HOT: august	0.859	0.856	0.856	0.858
HOT: september	1.28	1.27	1.27	1.27
HOT: october	1.33	1.33	1.33	1.33
HOT: november	1.25	1.25	1.25	1.25
HOT: december	1.23	1.23	1.23	1.23
HOT: tuesday	1.02	1.02	1.02	1.02
HOT: wednesday	1.03	1.03	1.03	1.03
HOT: thursday	0.898	0.896	0.899	0.898
HOT: friday	0.917	0.914	0.919	0.917
HOT: pm1530	0.947	0.946	0.945	0.946
HOT: pm1600	0.874	0.875	0.871	0.873
HOT: pm1630	0.988	0.990	0.984	0.987
HOT: pm1700	1.08	1.09	1.08	1.08
HOT: pm1730	1.23	1.24	1.23	1.23
HOT: pm1800	1.50	1.51	1.49	1.50
HOT: pm1830	1.41	1.42	1.41	1.41

Table 132: Model 14b AM - Five Income Groups - Odds Ratios

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	0.245	12.6	0.106	0.000	18046.40
$avgSpeed^2$	1.00	1.00	1.00	1.00	1.00
tollAmount	0.502	0.500	0.501	0.550	0.638
transponderCount	0.999	0.999	1.00	0.999	1.00
HOT: congested50	5.07	4.96	4.75	6.10	4.48
HOT: hhEdu	0.849	0.893	0.891	0.723	0.893
HOT: hhAge	0.971	0.997	0.907	0.924	0.602
HOT: income/hhSize	1.00	1.00	1.00	1.00	1.00
HOT: income	0.874	0.577	0.937	2.94	0.532
HOT: hhSize	0.929	1.05	1.01	1.00	0.836
HOT: segmentCount	2.63	2.52	2.60	2.32	1.95
HOT: february	1.36	1.29	1.31	1.19	1.16
HOT: march	1.25	1.28	1.26	1.24	1.28
HOT: april	1.48	1.47	1.48	1.27	1.38
HOT: may	1.45	1.46	1.53	1.38	1.36
HOT: june	1.28	1.37	1.31	1.38	1.19
HOT: july	1.11	1.22	1.13	1.19	1.23
HOT: august	1.50	1.55	1.54	1.38	1.37
HOT: september	1.75	1.84	1.87	1.79	1.45
HOT: october	1.72	1.85	1.85	1.77	1.56
HOT: november	1.44	1.67	1.63	1.63	1.44
HOT: december	1.16	1.27	1.33	1.40	1.16
HOT: tuesday	1.34	1.23	1.25	1.33	1.26
HOT: wednesday	1.35	1.23	1.27	1.27	1.25
HOT: thursday	1.33	1.24	1.22	1.30	1.22
HOT: friday	0.408	0.408	0.396	0.476	0.570
HOT: am630	5.39	4.86	5.02	5.21	3.07
HOT: am700	7.47	6.53	7.67	8.38	4.14
HOT: am730	8.57	6.48	8.42	6.84	4.72
HOT: am800	7.72	5.91	6.32	5.19	4.68
HOT: am830	6.11	4.53	4.84	5.11	2.30
HOT: am900	3.49	2.87	3.24	2.58	1.95
HOT: am930	1.54	1.39	1.47	1.13	1.12

Table 133: Model 14b PM - Five Income Groups - Odds Ratios

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	0.319	0.324	4.58	0.210	0.000
$avgSpeed^2$	1.00	1.00	1.00	1.00	1.00
tollAmount	0.650	0.627	0.606	0.647	0.766
transponderCount	1.00	1.00	1.00	1.00	1.00
HOT: congested40	3.92	3.83	4.15	3.94	3.57
HOT: hhEdu	0.819	0.826	0.735	0.722	0.919
HOT: hhAge	1.01	1.00	0.931	0.929	0.832
HOT: income/hhSize	1.00	1.00	1.00	1.00	1.00
HOT: income	0.895	0.894	0.712	1.13	4.08
HOT: hhSize	0.987	0.980	1.08	0.875	0.852
HOT: segmentCount	2.91	2.94	3.16	3.30	3.14
HOT: february	0.938	0.935	0.911	0.867	0.979
HOT: march	0.959	0.947	0.920	0.876	0.909
HOT: april	0.779	0.762	0.751	0.642	0.611
HOT: may	0.687	0.700	0.675	0.640	0.614
HOT: june	0.673	0.717	0.670	0.689	0.650
HOT: july	0.693	0.731	0.650	0.669	0.561
HOT: august	0.832	0.907	0.829	0.796	0.624
HOT: september	1.15	1.39	1.32	1.17	0.761
HOT: october	1.20	1.46	1.36	1.21	0.761
HOT: november	1.10	1.38	1.35	1.07	0.699
HOT: december	1.13	1.30	1.31	1.15	0.776
HOT: tuesday	1.02	1.01	1.02	0.959	1.04
HOT: wednesday	1.05	1.02	1.01	0.977	0.988
HOT: thursday	0.933	0.889	0.879	0.772	0.842
HOT: friday	0.969	0.887	0.890	0.844	1.09
HOT: pm1530	0.999	0.944	0.892	0.752	1.05
HOT: pm1600	0.897	0.856	0.880	0.831	0.912
HOT: pm1630	0.934	0.964	1.19	0.819	1.04
HOT: pm1700	1.02	1.09	1.25	0.791	1.08
HOT: pm1730	1.10	1.26	1.41	1.07	1.83
HOT: pm1800	1.33	1.52	1.74	1.40	2.23
HOT: pm1830	1.27	1.41	1.68	1.23	2.23

Table 134: Model 15 Odds Ratios

	AM Peak – Model 15	PM Peak – Model 15
Intercept	0.292	0.163
avgSpeed^2	1.00	1.00
Toll/segmentCount	0.0930	0.355
Transponder Count	0.998	1.00
HOT: congested50	6.15	
HOT: congested40		3.73
HOT: hhEdu	0.842	0.775
HOT: hhAge	0.976	0.998
HOT: log(hhIncomedollars)	1.22	1.22
HOT: hhSize	0.966	0.991
HOT: february	1.30	0.967
HOT: march	1.24	1.00
HOT: april	1.44	0.989
HOT: may	1.44	0.928
HOT: june	1.27	0.997
HOT: july	1.14	0.977
HOT: august	1.47	1.01
HOT: september	1.71	1.17
HOT: october	1.71	1.20
HOT: november	1.47	1.12
HOT: december	1.18	1.06
HOT: tuesday	1.28	1.04
HOT: wednesday	1.28	1.07
HOT: thursday	1.27	1.04
HOT: friday	0.437	1.07
HOT: am630	3.59	
HOT: am700	4.59	
HOT: am730	5.09	
HOT: am800	4.54	
HOT: am830	3.54	
HOT: am900	2.19	
HOT: am930	1.02	
HOT: pm1530		1.19
HOT: pm1600		1.26
HOT: pm1630		1.38
HOT: pm1700		1.41
HOT: pm1730		1.40
HOT: pm1800		1.47
HOT: pm1830		1.35

Table 135: Model 16 AM Peak Odds Ratios

	AM Peak – Model 16a	AM Peak – Model 16b	AM Peak – Model 16c
Intercept	1.38	0.156	4.25
$avgSpeed^2$	1.00	1.00	1.00
Toll/segmentCount	0.0600	0.0567	0.0569
Toll/log(income)	4.61	5.43	5.18
transponderCount	0.999	0.999	0.999
HOT: congested50	5.45	5.39	5.41
HOT: log(income)/hhSize	1.02	1.06	
HOT: income/hhSize			1.00
HOT: income		1.25	0.887
HOT: hhSize		1.00	1.06
HOT: hhAge	0.999	0.978	0.972
HOT: hhEdu	0.899	0.850	0.846
HOT: february	1.29	1.29	1.29
HOT: march	1.23	1.24	1.24
HOT: april	1.42	1.43	1.43
HOT: may	1.43	1.44	1.44
HOT: june	1.27	1.27	1.27
HOT: july	1.14	1.13	1.14
HOT: august	1.47	1.47	1.47
HOT: september	1.73	1.73	1.74
HOT: october	1.73	1.74	1.74
HOT: november	1.49	1.49	1.50
HOT: december	1.18	1.18	1.19
HOT: tuesday	1.27	1.28	1.28
HOT: wednesday	1.27	1.27	1.27
HOT: thursday	1.26	1.27	1.27
HOT: friday	0.443	0.444	0.443
HOT: am630	3.47	3.48	3.50
HOT: am700	4.65	4.69	4.71
HOT: am730	5.13	5.17	5.20
HOT: am800	4.49	4.53	4.55
HOT: am830	3.55	3.59	3.59
HOT: am900	2.21	2.23	2.22
HOT: am930	1.05	1.06	1.05

Table 136: Model 16 PM Peak Odds Ratios

	PM Peak – Model 16a	PM Peak – Model 16b	PM Peak – Model 16c
Intercept	2.17	0.0885	0.378
$avgSpeed^2$	1.00	1.00	1.00
Toll/segmentCount	0.00572	0.00481	0.00485
Toll/log(income)	50109	73953	72342
transponderCount	1.00	1.00	1.00
HOT: congested40	3.70	3.76	3.76
HOT: log(income)/hhSize	0.983	1.02	
HOT: income/hhSize			1.00
HOT: income		1.39	1.19
HOT: hhSize		0.993	1.02
HOT: hhAge	1.03	0.998	0.996
HOT: hhEdu	0.876	0.806	0.804
HOT: february	0.945	0.944	0.944
HOT: march	0.971	0.970	0.969
HOT: april	0.897	0.895	0.895
HOT: may	0.828	0.823	0.823
HOT: june	0.871	0.865	0.866
HOT: july	0.863	0.856	0.857
HOT: august	0.932	0.926	0.927
HOT: september	1.20	1.20	1.20
HOT: october	1.22	1.23	1.23
HOT: november	1.14	1.14	1.15
HOT: december	1.11	1.12	1.12
HOT: tuesday	1.02	1.02	1.02
HOT: wednesday	1.03	1.03	1.03
HOT: thursday	0.928	0.920	0.920
HOT: friday	0.947	0.938	0.937
HOT: pm1530	1.02	1.01	1.01
HOT: pm1600	0.970	0.963	0.965
HOT: pm1630	1.03	1.02	1.03
HOT: pm1700	1.07	1.07	1.07
HOT: pm1730	1.15	1.15	1.15
HOT: pm1800	1.34	1.34	1.35
HOT: pm1830	1.31	1.32	1.32

Table 137: Model 17 AM Peak Odds Ratios

	AM Peak – Model 17a	AM Peak – Model 17b	AM Peak – Model 17c
Intercept	4.06	0.233	2.34
$avgSpeed^2$	1.00	1.00	1.00
$tollAmount^2$		0.935	
Toll/segmentCount			0.0988
Toll/log(income)	0.000588		
transponderCount	0.999	0.999	0.999
HOT: congested50	4.93	3.84	4.55
HOT: income/hhSize	1.000	1.00	1.00
HOT: income	0.639	0.817	0.824
HOT: hhSize	1.06	1.05	1.06
HOT: hhAge	0.970	0.971	0.971
HOT: hhEdu	0.872	0.873	0.869
HOT: segmentCount	2.52	2.14	1.42
HOT: february	1.30	1.25	1.28
HOT: march	1.26	1.21	1.24
HOT: april	1.46	1.34	1.43
HOT: may	1.46	1.37	1.44
HOT: june	1.31	1.27	1.28
HOT: july	1.16	1.13	1.14
HOT: august	1.51	1.40	1.49
HOT: september	1.79	1.69	1.78
HOT: october	1.78	1.68	1.79
HOT: november	1.56	1.50	1.55
HOT: december	1.24	1.19	1.21
HOT: tuesday	1.27	1.27	1.27
HOT: wednesday	1.27	1.28	1.27
HOT: thursday	1.27	1.27	1.26
HOT: friday	0.417	0.538	0.440
HOT: am630	4.85	3.48	3.82
HOT: am700	6.82	4.89	5.57
HOT: am730	7.21	5.14	6.00
HOT: am800	6.23	4.55	5.11
HOT: am830	4.87	3.86	4.07
HOT: am900	3.04	3.05	2.53
HOT: am930	1.44	2.20	1.23

Table 138: Model 17 PM Peak Odds Ratios

	PM Peak – Model 17a	PM Peak – Model 17b	PM Peak – Model 17c
Intercept	0.390	0.178	0.335
$avgSpeed^2$	1.00	1.00	1.00
$tollAmount^2$		0.954	
Toll/segmentCount			0.204
Toll/log(income)	0.00650		
transponderCount	1.00	1.00	1.00
HOT: congested40	3.91	3.43	3.77
HOT: income/hhSize	1.00	1.00	1.00
HOT: income	0.859	0.951	0.947
HOT: hhSize	1.03	1.03	1.03
HOT: hhAge	0.993	0.994	0.994
HOT: hhEdu	0.809	0.811	0.816
HOT: segmentCount	2.99	2.59	2.38
HOT: february	0.931	0.917	0.926
HOT: march	0.943	0.897	0.932
HOT: april	0.759	0.736	0.769
HOT: may	0.687	0.676	0.697
HOT: june	0.689	0.688	0.705
HOT: july	0.698	0.690	0.709
HOT: august	0.856	0.764	0.835
HOT: september	1.27	1.05	1.21
HOT: october	1.32	1.08	1.25
HOT: november	1.25	1.05	1.18
HOT: december	1.22	1.04	1.17
HOT: tuesday	1.02	1.00	1.01
HOT: wednesday	1.03	0.996	1.01
HOT: thursday	0.896	0.833	0.864
HOT: friday	0.915	0.844	0.878
HOT: pm1530	0.945	0.868	0.915
HOT: pm1600	0.870	0.739	0.818
HOT: pm1630	0.982	0.768	0.893
HOT: pm1700	1.08	0.810	0.962
HOT: pm1730	1.22	0.908	1.09
HOT: pm1800	1.49	1.13	1.36
HOT: pm1830	1.41	1.16	1.33

Table 139: Mixed Logit Model 1a – AM Peak – 5 Income Groups Odds Ratios

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	0.0114	0.329	0.000391	25.3	0.000
$avgSpeed^2$	1.00	0.999	0.999	0.999	1.00
tollAmount	0.442	0.471	0.475	0.730	0.846
transponderCount	1.00	1.00	1.00	1.00	1.00
HOT: congested50	6.93	7.18	8.36	12.9	9.49
HOT: hhEdu	1.03	0.966	0.848	0.973	1.20
HOT: hhAge	0.972	1.03	0.912	0.948	0.829
HOT: income/hhSize	1.00	1.00	1.00	1.00	1.00
HOT: income	1.12	0.13	1.66	0.543	12.5
HOT: hhSize	0.960	0.984	0.972	1.10	0.748
HOT: segmentCount	2.80	2.82	2.86	2.81	2.43
HOT: february	1.58	1.52	1.64	1.36	1.31
HOT: march	1.59	1.64	1.58	1.50	1.79
HOT: april	1.96	1.94	1.98	1.60	1.98
HOT: may	1.69	1.66	1.94	1.78	1.70
HOT: june	1.71	1.69	1.83	2.05	1.55
HOT: july	1.36	1.40	1.48	1.53	1.29
HOT: august	1.87	1.82	2.05	1.58	1.30
HOT: september	2.02	1.91	2.05	2.06	1.56
HOT: october	2.16	2.04	2.16	2.00	1.76
HOT: november	1.58	1.62	1.64	1.63	1.52
HOT: december	1.05	1.07	1.16	1.12	0.953
HOT: tuesday	1.48	1.39	1.35	1.43	1.61
HOT: wednesday	1.62	1.45	1.46	1.59	1.75
HOT: thursday	1.51	1.45	1.42	1.61	1.73
HOT: friday	0.450	0.446	0.400	0.444	0.706
HOT: am630	5.83	5.39	6.16	0.749	2.49
HOT: am700	6.90	6.43	8.36	9.89	4.19
HOT: am730	6.74	5.22	6.50	6.82	3.16
HOT: am800	4.30	3.52	3.73	4.03	2.31
HOT: am830	2.76	2.30	2.29	2.92	1.22
HOT: am900	1.45	1.15	1.23	1.47	0.655
HOT: am930	0.617	0.555	0.579	0.656	0.411

Table 140: Mixed Logit Model 1a – PM Peak – 5 Income Groups Odds Ratios

	Segment A $0-50k	Segment B $50-100k	Segment C $100-150k	Segment D $150-200k	Segment E $200k+
Intercept	0.0154	0.0638	49.5	0.127	0.000
$avgSpeed^2$	1.00	1.00	1.00	1.00	1.00
tollAmount	0.698	0.749	0.744	1.03	1.675
transponderCount	1.01	1.01	1.01	1.01	1.01
HOT: congested40	5.52	5.40	5.36	4.68	5.16
HOT: hhEdu	1.02	0.969	1.02	0.843	0.903
HOT: hhAge	1.00	0.999	1.01	0.891	0.981
HOT: income/hhSize	1.00	1.00	1.00	1.00	1.00
HOT: income	0.956	0.827	0.454	0.958	2.23
HOT: hhSize	1.01	1.03	1.04	0.918	0.965
HOT: segmentCount	4.53	4.89	5.35	5.00	4.94
HOT: february	0.904	0.918	0.857	0.909	0.831
HOT: march	0.916	0.917	0.864	0.994	0.864
HOT: april	0.759	0.753	0.692	0.688	0.586
HOT: may	0.623	0.658	0.602	0.722	0.594
HOT: june	0.624	0.647	0.585	0.738	0.479
HOT: july	0.627	0.658	0.576	0.732	0.511
HOT: august	0.808	0.883	0.780	0.937	0.528
HOT: september	1.27	1.53	1.30	1.42	0.702
HOT: october	1.36	1.70	1.42	1.33	0.841
HOT: november	1.19	1.47	1.37	1.25	0.662
HOT: december	1.10	1.30	1.33	1.16	0.764
HOT: tuesday	0.988	1.01	0.947	0.993	1.03
HOT: wednesday	1.02	1.02	0.987	1.03	0.977
HOT: thursday	0.877	0.865	0.828	0.774	0.821
HOT: friday	0.930	0.866	0.889	0.856	1.19
HOT: pm1530	0.916	0.887	0.837	0.686	1.06
HOT: pm1600	0.852	0.823	0.887	0.811	0.983
HOT: pm1630	1.04	1.04	1.21	1.05	1.23
HOT: pm1700	1.23	1.34	1.50	1.24	1.24
HOT: pm1730	1.49	1.66	1.89	1.78	2.14
HOT: pm1800	1.91	1.92	2.21	1.88	2.03
HOT: pm1830	1.68	1.73	1.88	1.44	2.10

REFERENCES

Bhat, C. R., Castelar, S. (2002). A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. *Transportation Research Part B: Methodological*, 36(7), 593-616.

Börjesson, M. (2008). Joint RP–SP data in a mixed logit analysis of trip timing decisions. *Transportation Research Part E: Logistics and Transportation Review*, 44(6), 1025-1038.

Boyles, S. D., Kockelman, K. M., Waller, S. T. (2010). Congestion pricing under operational, supply-side uncertainty. *Transportation Research Part C: Emerging Technologies*, 18(4), 519-535.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L., Cutler, A. Random Forests. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/

Brownstone, D., Small, K. A. (2005). Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A: Policy and Practice*, 39(4), 279-293.

Brownstone, D., Ghosh, A., Golob, T. F., Kazimi, C., Van Amelsfort, D. (2003). Drivers' willingness-to-pay to reduce travel time: evidence from the San Diego I-15 congestion pricing project. *Transportation Research Part A: Policy and Practice*, 37(4), 373-387.

Burris, M. W., Alemazkoor, N., Benz, R., Wood, N. S. (2014). The impact of HOT lanes on carpools. *Research in Transportation Economics*, 44, 43-51.

Burris, M. W., Figueroa, C. F. (2010). Analysis of traveler characteristics by mode choice in HOT corridors. *Journal of the Transportation Research Forum*, 45(2), 103-117.

Burris, M. W., Patil, S. (2009). Estimating the Benefits of Managed Lanes (No. UTCM 08-05-04).

Burris, M. W., Pendyala, R. M. (2002). Discrete choice models of traveler participation in differential time of day pricing programs. *Transport Policy*, 9(3), 241-251.

Burris, M. W., Nelson, S., Kelly, P., Gupta, P., Cho, Y. (2012). Willingness to Pay for High-Occupancy Toll Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 2297, 47-55.

Calfee, J., Winston, C. (1998). The value of automobile travel time: implications for congestion policy. *Journal of Public Economics*, 69(1), 83-102.

Calfee, J., Winston, C., Stempski, R. (2001). Econometric issues in estimating consumer preferences from stated preference data: a case study of the value of automobile travel time. *Review of Economics and Statistics*, 83(4), 699-707.

Carrion, C., Levinson, D. (2012). Value of travel time reliability: A review of current evidence. *Transportation Research Part A: Policy and Practice*, 46(4), 720-741.

Carrion, C., Levinson, D. (2013). Valuation of travel time reliability from a GPS-based experimental design. *Transportation Research Part C: Emerging Technologies*, 35, 305-323.

Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks. *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, 697-711.

Cherchi, E., Cirillo, C. (2007). A mixed logit mode choice model on panel data: accounting for systematic and random variations on responses and preferences. *Transportation Research Board Annual Meeting*.

Chin, K.-K. (2010). The Singapore Experience: The evolution of technologies, costs and benefits, and lessons learnt. *Joint Transport Research Centre Conference Proceedings*.

de Palma, A., Lindsey, R. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C: Emerging Technologies*, 19(6), 1377-1399.

Devarasetty, P. C., Burris, M. W., Douglass Shaw, W. (2012). The value of travel time and reliability-evidence from a stated preference survey and actual usage. *Transportation Research Part A: Policy and Practice*, 46(8), 1227-1240.

Devarasetty, P., Burris, M. W., Huang, C. (2013). Examining the Differences Between Travelers' Revealed Versus Actual Travel Time Savings. *Transportation Research Board Annual Meeting*.

Dinsdale, E. (2013). Random Forests. *Metagenomics. Statistics*. Retrieved from https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html.

Efron, B., Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Elango, V. V., Guensler, R. (2014). Collection, Screening, and Evaluation of Vehicle Occupancy Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2470(1), 142-151.

Epsilon Targeting. (2013). List Hygiene & Data Enhancement Report..

Evans, I. V., Bhatt, K. U., Turnbull, K. F. (2003). Road Value Pricing-Traveler Response to Transportation System Changes. *Transit Cooperative Research Program*.

Federal Highway Administration. (2006). *Travel Time Reliability: Making It There on Time, All the Time* (No. FHWA-HOP-06-070).

Federal Highway Administration. (2014). High Occupancy Toll (HOT) Lanes Marketing Toolkit. Retrieved from http://ops.fhwa.dot.gov/publications/fhwahop12031/index.htm

Fosgerau, M., Engelson, L. (2011). The value of travel time variance. *Transportation Research Part B: Methodological*, 45(1), 1-8.

Georgia Department of Transportation. (2015). "Express Lanes." Retrieved from http://www.dot.ga.gov/DS/GEL.

Golob, J. M., Golob, T. F. (2001). Studying Road Pricing Policy with Panel Data Analysis: The San Diego I-15 HOT Lanes. In David A. Hensher (Ed.), *Travel Behaviour Research: The Leading Edge* (869-883). Elsevier.

Goodall, N., Smith, B. L. (2010). What Drives Decisions of Single-Occupant Travelers in High-Occupancy Vehicle Lanes?. *Transportation Research Record: Journal of the Transportation Research Board*, 2178(1), 156-161.

Goodwin, P. B. (1992). A review of new demand elasticities with special reference to short and long run effects of price changes. *Journal of Transport Economics and Policy*, 26(2), 155-169.

Graham, D. J., Glaister, S. (2004). Road traffic demand elasticity estimates: a review. *Transport Reviews*, 24(3), 261-274.

Guensler, R., Elango, V., Guin, A., Hunter, M., Laval, J., Araque, S., Box, S., Colberg, K., Castrillon, F., D'Ambrosio, K., Duarte, D., Kamiya, K., Khoeini, S., Palinginis, E., Peesapati, L., Rome, C., Sheikh, A., Smith, K., Toth, C., Vo, T., Zinner, S. (2013). Atlanta I-85 HOV-to-HOT Conversion: Analysis of Vehicle and Person Throughput. Prepared for the Georgia Department of Transportation, Atlanta, GA. Georgia Institute of Technology. Atlanta, GA. October 2013.

Guin, A., Hunter, M., Guensler, R. (2008). Analysis of Reduction in Effective Capacities of High-Occupancy Vehicle Lanes Related to Traffic Behavior. *Transportation Research Record: Journal of the Transportation Research Board* 2065, 47-53.

Guo, X. Yang, H. (2009). Pareto-improving congestion pricing and revenue refunding with multiple user classes. *Transportation Research Part B: Methodological*, 44(8), 972-982.

Hamilton, C. (2010). Revisiting the Cost of the Stockholm Congestion Charging System. *Joint Transport Research Center. Conference Proceedings*.

Han, C. P., Li, J. (2009). Evaluating Estimation Techniques of Transportation Price Elasticity. *Transportation Research Record: Journal of the Transportation Research Board*, 2115(1), 94-101.

He, X., Liu, H. X., Cao, X. J. (2012). Estimating value of travel time and value of reliability using dynamic toll data. *Transportation Research Board Annual Meeting*.

Henningsen, A., Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. Computational Statistics 26(3), 443-458. DOI 10.1007/s00180-010-0217-1.

Hess, S., Bierlaire, M., Polak, J. (2005). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39(2), 221-236.

Hirschman, I., McKnight, C., Pucher, J., Paaswell, R. E., Berechman, J. (1995). Bridge and tunnel toll elasticities in New York. *Transportation*, 22(2), 97-113.

Hlavac, M. (2014). stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables. R package version 5.1. http://CRAN.R-project.org/package=stargazer

HOV Strategic Implementation Plan Atlanta Region. (2003). "http://www.dot.ga.gov/Projects/studies/Pages/HOV.aspx".

Janson, M., Levinson, D. (2014). HOT or not: Driver elasticity to price on the MnPASS HOT lanes. *Research in Transportation Economics*, 44, 21-32.

Jiang, M., Morikawa, T. (2004). Theoretical analysis on the variation of value of travel time savings. *Transportation Research Part A: Policy and Practice*, 38(8), 551-571.

Khoeini, S., Guensler, R. (2013). Pricing Impact on Users: Socioeconomic Study on I-85 HOV2 to HOT3 Conversion. Poster presented at 2013 TRB Freeway and Managed Lanes Operations Meeting and Conference, Atlanta, GA.

Khoeini, S., Xu, Y., Guensler, R. (2013). Marketing Data Application in Transportation Studies: Household Level Accuracy and Inclusiveness Evaluation. Working Paper.

Khoeini, S. (2014). *Modeling Framework for Socio-Economic Analysis of Managed Lanes* (Doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.

Khoeini, S., Guensler, R. (2014). Using vehicle value as a proxy for income: A case study on Atlanta's I-85 HOT lane. *Research in Transportation Economics*, 44, 33-42.

King, D., Manville, M., Shoup, D. (2007). The political calculus of congestion pricing. *Transport Policy*, 14(2), 111-123.

Lam, T. C., Small, K. A. (2001). The value of time and reliability: measurement from a value pricing experiment. *Transportation Research Part E: Logistics and Transportation Review*, 37(2), 231-251.

Li, J. (2001). Explaining High Occupancy Toll Lane Use. *Transportation Research Part D: Transport and Environment*, 6, 61-74.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.

Lindsey, R. (2009). Cost recovery from congestion tolls with random capacity and demand. *Journal of Urban Economics*, 66(1), 16-24.

Litman, T. (2006). London Congestion Pricing: Implications for Other Cities. Victoria Transport Policy Institute.

Litman, T. (2004). Transit price elasticities and cross-elasticities. *Journal of Public Transportation*, 7, 37-58.

Litman, T. (2007). Transportation elasticities: How Prices and Other Factors Affect Travel Behavior. Victoria Transport Policy Institute.

Litman, T. (2010). Changing vehicle travel price sensitivities: the rebounding rebound effect. Victoria Transport Policy Institute.

Liu, HX., He, X., Recker, W. (2007). Estimation of the time-dependency of values of travel time and its reliability from loop detector data. *Transportation Research Part B: Methodological*, 41(4), 448-461.

Margiotta, R., Lomax, T., Hallenbeck, M., Dowling, R., Skabardonis, A., Turner, S. (2013). Analytical Procedures for Determining the Impacts of Reliability Mitigation Strategies. Publication S2-L03-RR-1. Strategic Highway Research Program, Transportation Research Board.

McFadden, D., Train, K. (2000). Mixed MNL models for discrete response. *Journal of applied Econometrics*, 15(5), 447-470.

Milborrow, S. (2011). rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart. R Package.

Mohring, H., Schroeter, J., Wiboonchutikula, P. (1987). The values of waiting time, travel time, and a seat on a bus. *The RAND Journal of Economics*, 40-56.

Nakamura, K., Kockelman, K. M. (2002). Congestion pricing and roadspace rationing: an application to the San Francisco Bay Bridge corridor. *Transportation Research Part A: Policy and Practice*, 36(5), 403-417.

Nelson, J.I., Guensler, R., Li, H. (2008). Geographic and Demographic Profiles of Morning Peak-Hour Commuters on Highways in North Metropolitan Atlanta, Georgia. *Transportation Research Record: Journal of the Transportation Research Board*, 2067, 26–37.

Oum, T. H. (1989). Alternative demand models and their elasticity estimates. *Journal of Transport Economics and Policy*, 163-187.

Pahaut, S., Sikow, C. (2006). History of thought and prospects for road pricing. *Transport Policy*, 13(2), 173-176.

Papola, A. (2004). Some developments on the cross-nested logit model. *Transportation Research Part B: Methodological*, 38(9), 833-851.

Patterson, T., Levinson, D. M. (2008). Lexus lanes or corolla lanes? Spatial use and equity patterns on the I-394 MnPASS lanes. University of Minnesota: Nexus Research Group.

Pierce, G., Shoup, D. (2013). Getting the prices right: an evaluation of pricing parking by demand in San Francisco. *Journal of the American Planning Association*, 79(1), 67-81.

Pigou, A. C. (1920). The economics of welfare. Palgrave Macmillan.

Pozdena, R. (2010). Improving Highway Efficiency and Investment Policy Through Pricing: A Primer. Puget Sound Regional Council.

Pratt, R. H. (2003). TCRP Research Results Digest, No. 61: Traveler Response to Transportation System Changes: An Interim Introduction to the Handbook. Transportation Research Board of the National Academies, Washington, D.C.

Rouwendal, J., Verhoef, E. T. (2006). Basic economic principles of road pricing: From theory to applications. *Transport policy*, 13(2), 106-114.

Santos, G., Behrendt, H., Maconi, L., Shirvani, T., Teytelboym, A. (2010). Part I: Externalities and economic policies in road transport. *Research in Transportation Economics*, 28(1), 2-45.

Schrank, D., Eisele, B., Lomax, T. (2012). TTI's 2012 Urban Mobility Report. Texas A&M Transportation Institute. The Texas A&M University System.

Shalizi, C. (2013). Classification and Regression Trees [PDF Document]. Retrieved from http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

Sheikh, A., Guin, A., Guensler, R. (2014). Value of Travel Time Savings: Evidence from I-85 Express Lanes in Atlanta, Georgia. *Transportation Research Record: Journal of the Transportation Research Board*, 2470, 161–168.

Sheikh, A., Misra, A., Guensler, R. (2016). High Occupancy Toll Lane Decision Making: Income Effects on Atlanta's I-85 Express Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 2531.

Simmons, A. (October 14, 2013). "HOT Lane Use Grows as Drivers Grumble." *Atlanta Journal-Constitution*. Retrieved from http://www.myajc.com/news/news/hot-lane-use-grows-as-drivers-grumble/nbNjz/

Small, K. A., Winston, C., Yan, J. (2005). *Uncovering the distribution of motorists' preferences for travel time and reliability*. Econometrica, 73(4), 1367-1382.

Smith, K. S. (2011). A profile of HOV lane vehicle characteristics on I-85 prior to HOV-to-HOT conversion. (Master's thesis). Georgia Institute of Technology, Atlanta, GA.

Sullivan, E. C. (2002). State Route 91 value-priced express lanes: updated observations. *Transportation Research Record: Journal of the Transportation Research Board*, 1812(1), 37-42.

Szumilas, M. (2010). Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry, 227–9.

Evans, J. E., Bhatt, K. U., Turnbull, K. F. (2003). TCRP Report 95: Road Value Pricing – Traveler Response to Transportation System Changes. Transportation Research Board of the National Academies, Washington, D.C.

Train, K. E. (1986). *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. Cambridge, MA: MIT press.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge, England: Cambridge University Press.

United States Census Bureau / American FactFinder. (2015). 2009 – 2013 American Community Survey 5-Year Estimates. Retrieved from http://factfinder2.census.gov.

United States Census Bureau. (2014). TIGER/Line Shapefiles Technical Documentation. Retrieved from http://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2014/TGRSHP2014_TechDoc.pdf.

U.S. Bureau of Labor Statistics. (2012). Metropolitan and Nonmetropolitan Area Occupational Employment and Wage Estimates. Retrieved from http://bls.gov/oes/current/oes_12060.htm#00-0000.

Vovsha, P., Bekhor, S. (1998). Link-nested logit model of route choice: overcoming route overlapping problem. *Transportation Research Record: Journal of the Transportation Research Board*, 1645(1), 133-142.

Washington, S., Wolf, J., Guensler, R. (1997). Binary recursive partitioning method for modeling hot-stabilized emissions from motor vehicles. Transportation Research Record: Journal of the Transportation Research Board, (1587), 96-105.

Wood, N., Burris, M., Danda, S. (2014). Examination of Paid Travel on I-85 Express Lanes. Transportation Research Record: Journal of the Transportation Research Board, (2450), 44-51.

Yan, J., Small, K. A., Sullivan, E. C. (2002). Choice models of route, occupancy, and time of day with value-priced tolls. *Transportation Research Record: Journal of the Transportation Research Board*, 1812(1), 69-77.